# Perceptual Mixing for Musical Production

**Michael John Terrell**

Submitted to the University of London in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Queen Mary University of London

2012

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline. I acknowledge the helpful guidance and support of my supervisor, Mark Sandler.

# Abstract

A general model of music mixing is developed, which enables a mix to be evaluated as a set of acoustic signals. A second model describes the mixing process as an optimisation problem, in which the errors are evaluated by comparing sound features of a mix with those of a reference mix, and the parameters are the controls on the mixing console. Initial focus is placed on live mixing, where the practical issues of: live acoustic sources, multiple listeners, and acoustic feedback, increase the technical burden on the mixing engineer. Using the two models, a system is demonstrated that takes as input reference mixes, and automatically sets the controls on the mixing console to recreate their objective, acoustic sound features for all listeners, taking into account the practical issues outlined above. This reduces the complexity of mixing live music to that of recorded music, and unifies future mixing research.

Sound features evaluated from audio signals are shown to be unsuitable for describing a mix, because they do not incorporate the effects of listening conditions, or masking interactions between sounds. Psychophysical test methods are employed to develop a new perceptual sound feature, termed the loudness balance, which is the first loudness feature to be validated for musical sounds. A novel, perceptual mixing system is designed, which allows users to directly control the loudness balance of the sounds they are mixing, for both live and recorded music, and which can be extended to incorporate other perceptual features. The perceptual mixer is also employed as an analytical tool, to allow direct measurement of mixing best practice, to provide fully-automatic mixing functionality, and is shown to be an improvement over current heuristic models. Based on the conclusions of the work, a framework for future automatic mixing is provided, centred on perceptual sound features that are validated using psychophysical methods.

*For Kenny and Bettie*

# Acknowledgements

I would like to take this opportunity to thank the EPSRC, and everyone in the Centre for Digital Music at Queen Mary University for making it possible for me to complete this Ph.D.

I would like to thank my supervisor, Mark Sandler, for taking me on as a student in less than ideal circumstances, for offering the big picture view whenever it was needed, and most of all, for trusting me enough to let me get on with it! I would also like to single out two other colleagues whose help has been invaluable; Enrique Perez-Gonzalez and Andy Simpson. Without Enrique's (sometimes blunt!) advice on audio engineering practice I would have been truly lost in my first year. I cannot overstate how much I learned from Enrique, and to put it in his words, his help was "pretty ok". In the latter part of this Ph.D it has been the constant discussions with Andy that have driven the work forward. His seemingly innocuous suggestion that I look at loudness models was the root of our working relationship and friendship, which I hope will continue long into the future. Thanks also to all the great friends I've made here in C4DM; the past 4 years amongst the most enjoyable of my life.

To my parents, thank you, as always, for the unwavering support throughout my life. I doubt I would've even started this Ph.D without your safety net beneath me (and specific thanks to my dad for the detailed proof read of this thesis!). Finally, I would like to thank my fiancee Jo, who met me during my transition from big bucks hedge fund to poor postgrad student. Thank you for supporting me, for being so understanding, for not minding having a geek for a boyfriend, and for feigning interest every time I asked you to look at yet another graph. I love you baby x.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Music Production, Automatic Mixing and Sound Features

---

Music production includes all stages in the creation of a piece of music, from the initial conception and composition, to the mastering and distribution of the finished product. Mixing is one part of the production process, within which the sounds from different sources, typically musical instruments, are combined to form a coherent piece of music, referred to as the 'mix'. It can be split into two broad classes, which are live mixing, and the mixing of recorded music. The rationale behind the two is the same, i.e. to balance the sounds to provide a good mix for the listener, but live mixing is more complex because there are additional practical issues to be taken into account, including the presence of live acoustic sources, and the need to provide mixes to numerous audience locations as well as to the performers themselves. Mixing of both live, and recorded music, is the subject of this thesis.

The system used to mix musical sounds is called a mixing console, and the signal processing tools within it are broadly termed audio effects. Audio effects operate on audio signals that are either generated by live instruments (and passed into the console using microphones), or that have been pre-recorded, and stored (typically as digital audio signals in modern day consoles). Advances in digital mixing technology have been significant in recent years, in part due to the rapid increase in computational power. One effect of this is that music production has become far more accessible. Amateur music producers, particularly musicians, are now able to do all stages of the production process in the comfort of their bedrooms, and can release their own

work without the need for record label backing. A second effect of these advances is an increase in the complexity of the mixing tools, which places a greater technical burden on amateur and professional mixing engineers alike.

Advances in music production technology have led to a recent surge of interest in the field of research known as automatic mixing. The use of the term 'automatic' to describe this field of research is potentially misleading, because some work merely 'assists' the mixing process. However, its use persists and is therefore retained here.[1] The main focus of automatic mixing work is non-expert practitioners within the music production community. Its aim is to enable them to produce music to a high standard, without requiring the technical expertise of a professional mixing engineer. New technology should enable non-experts to mix live music events, in which the practical issues are accounted for by the automatic system; and should help musicians to produce good quality mixes of their work, without the need for them to delve too deeply into the technical complexities of mixing. Furthermore, for professional applications, the automatic mixing tools should remove some of the functional burden from the production process, or should assist in the engineer's decision making process, enabling him to concentrate on the more creative aspects of his work.

In the next section, state of the art automatic mixing research is discussed. A common theme within all such work is the use of sound features to describe the objectives of the mixing task. Therefore, following the automatic mixing literature, sound features are discussed in general, including current auditory research that allows perceptual features, such as loudness, to be modelled. Finally, the objectives and an overview of this thesis are provided.

## 1.1 Automatic mixing

There are diverse approaches to automatic mixing, but in all cases, *sound features* are used to describe the objectives of the mixing task, and these are mapped to the *control parameters* on an audio effect. Existing automatic tools are classified into three groups: real-time automatic mixing tools, offline automatic mixing tools, and re-parameterised audio effects. Real-time automatic mixing tools continually evaluate the sound features of a set of audio signals, and provide the corresponding control data. They are classified as adaptive digital audio effects (A-DAFx) [Verfaille et al., 2006], because the control parameters on the audio effect change depending on

[1] The term 'automatic' stems from the early pioneering work by Perez-Gonzalez and Reiss [Perez-Gonzales and Reiss, 2007, 2008, 2009a,b, 2010], which provided fully-automatic mixing functionality.

features of the audio signal, i.e. they adapt. Off-line automatic mixing tools analyse the sound features of a predefined set of audio signals and derive a single, static set of control parameters to drive the associated audio effects. Re-parameterised audio effects are controlled using intuitive, high-level parameters, which are mapped to the standard low-level parameters of the effect, providing simpler control for non-expert users.

### 1.1.1 Real-time automatic mixing tools

The first real-time automatic mixing tools were developed by Dugan [1975, 1989], and Julstrom and Tichy [1976] for conference applications. The objective of these tools was to prevent amplification of inactive microphones, which in turn reduced the likelihood of acoustic feedback. Dugan used a microphone placed in the conference space to gain an estimate of the background noise level. Noise gates[2] were used on each microphone signal, and the threshold on each gate was set using the estimated background noise level. Julstrom used front and rear facing microphones at each speaker location. The direction of the sound source was estimated by comparing front and rear signals, and signals identified as coming from the rear (i.e. not from the direction of the speaker) were attenuated.

Real-time automatic mixing tools for live music were presented by Perez-Gonzales and Reiss [2007, 2008, 2009a,b, 2010]. Each tool focused on a specific mixing task. The automatic gain normalizer [Perez-Gonzales and Reiss, 2008] set the input gain on the mixing desk. The objective was to maximize the level to give good dynamic range while preventing distortion and clipping. The automatic panner [Perez-Gonzales and Reiss, 2007, 2010], panned[3] audio signals based on spectral content. Its objective was to reduce masking by spatially separating signals with similar frequency content. Additional subjective considerations were used to make the resultant panning more representative of a human mixing engineer, for example low frequency sources and vocals were not panned. Through subjective evaluation it was shown that the panning settings obtained from the automatic panner were equally good when compared to those set by an experienced engineer. The automatic fader and equaliser effects [Perez-Gonzales and Reiss, 2009a,b] set fader levels and a filter bank to ensure that the probable loudness of all audio signals was equal. When combined with the automatic gain normaliser and the automatic panner, the result was a fully-automatic mixing system. While this generality is very useful if no knowledge of the music

[2] A noise gate is an audio effect that attenuates audio signals whose level is below the threshold.

[3] Panning is the placement of a sound into either the left or the right loudspeaker in a stereo mix.

is available, it does not permit specification by the artist, as to how the mixes should sound. In addition, the objective of their mix - to give all audio signals equal loudness probability - was defined heuristically, and is yet to be fully validated. Perez-Gonzalez and Reiss ignored the contribution of the live acoustic sources to the mix, and acoustic-environmental effects, so application of their work was restricted to front of house[4] (FOH) mixes in large venues, where these assumptions are thought to be valid. However, their work has since been developed by Mansbridge et al. [2012] into a tool that mixes recorded music.

### 1.1.2 Offline automatic mixing tools

Off-line automatic mixing tools for recorded music have been presented by Kolasinski [2008] and Barchiesi and Reiss [2009, 2010]. They used features extracted from a mix, to determine the control parameters of audio effects that had been applied to its component audio signals. Kolasinski [2008] estimated the gain applied to each signal using the spectral histogram of the mix. This resulted in a non-unique solution so a heuristic optimisation algorithm was used. The problem of non-uniqueness was exacerbated when the number of tracks was increased. Barchiesi and Reiss [2009, 2010] sought the control parameters of equalisation filters and delays applied to each track, as well as the gains. Rather than using higher level features of the mix, the full time domain mix signal was modelled as a linear combination of the individual tracks, which had been convolved with linear time invariant (LTI) impulse responses. The impulse responses were then extracted on a frame-by-frame basis using the least squares method. If the equalisation applied to the tracks was LTI, i.e. performed using a finite impulse response (FIR) filter, and the estimation order of the filter was at least as long as the equalisation filter, then the exact mixing parameters could be found. For this reason their work was termed, “reverse engineering of a mix”. In their work they also identified the non-linear gain envelopes that had been applied to the tracks using a compressor. This was done by modelling the envelope across small frames of audio using a polynomial function. Due to the inherent non-linear nature and variability in compressor implementation it was not possible to relate the gain envelopes directly to compressor parameters.

An off-line automatic monitor mixing[5] technique for live music production was presented by Terrell and Reiss [2009]. In their model, the root-mean-square (RMS) sound level of each instrument at each performer location was evaluated by combining the sound from the live acoustic

[4]The front of house mix is the mix heard by the audience

[5]The monitor mixes are the mixes heard by the performers.

sources with the reinforced sound from the loudspeakers, where the level of the reinforced sound was controlled using gain. The sound was combined independent of frequency assuming incoherence between signals, and the mix was described using the relative level of its component instruments. The authors used a gradient descent method to find the gain controls for each instrument that gave a best fit to the subjectively defined reference mixes. The importance of each performer's reference mix was weighted to reflect the fact that the quality needed in a monitor mix can differ from one performer to the next. In their case study, the order of importance was rated to be: vocalist, guitarist, bassist then drummer. The optimisation algorithm was constrained to keep the feedback loop gain below -3 dB, and to keep the overall RMS level of the monitor mixes between 95 dBSPL and 105 dBSPL at each receiver location. They showed that the feedback constraint severely restricted the feasible solution space, which made it impossible to recreate the reference mixes, particularly at the drummer location. The interaction between acoustic signals generated by the same instrument but emitted from different sound sources was not included in the model, nor was the interaction with early room reflections, which are a significant simplification.

### 1.1.3 Re-parameterised audio effects

Another body of work which relates to automatic mixing is the re-parameterisation of audio processing devices, typically a standard audio effect, to operate using high-level controls, which in some cases are based on perceptual features. Reed [2000] presented a system that learned the equaliser adjustments needed to affect a certain type of perceptual change. His system was trained by asking users to make specific perceptual changes to training audio. Once trained the user could for example, "make the song sound brighter," using a single control. A similar approach was taken by Rafii and Pardo [2009], and Sabin and Pardo [2009a,b], in the training of reverberation and equalisation effects respectively. Terrell et al. [2010] presented an algorithm that automatically set the attack, release, hold and threshold parameters of a noise gate applied to a kick drum recording containing bleed from secondary sources. The objective was to minimise the distortion of the kick drum while ensuring a predefined reduction in the bleed level. Once the parameters had been set the gain control could be used to control the perceived 'strength' of the noise gate.

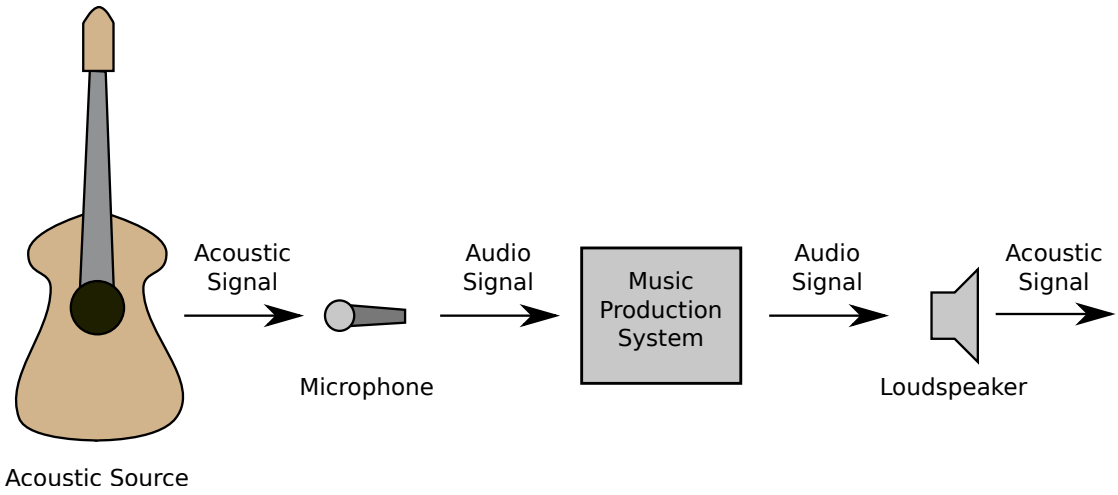


Figure 1.1: The different types of signal from which sound features are extracted. The acoustic source produces an acoustic signal (pressure fluctuations), which is converted into an audio signal (an electrical representation of the pressure fluctuations) by the microphone. The audio signal is processed by the music production system, and is sent to the loudspeaker, where it is converted back into an acoustic signal.

## 1.2 Sound features

A fundamental part of all automatic mixing applications is the use of sound features to describe their objective. These features can either be objective, i.e. mathematically defined signal properties; or perceptual, i.e. evaluated using a model that reflects human perception of sound. The sound features can be extracted from different points in the music production signal chain, which includes both acoustic and audio signals, as shown in Figure 1.1. The two types of signal differ in their units and scaling. Acoustic signals capture the absolute pressure variations in a sound and have units of Pascals (Pa), whereas audio signals are unscaled, electrical representations of the sound, and have units of Volts (V). The critical difference is that an acoustic signal contains information on the absolute level of the sound it represents, whereas an audio signal does not. This section contains a discussion of objective and perceptual sound features, and makes distinctions between those evaluated using acoustic or audio signals.

### 1.2.1 Objective sound features

Objective sound features are mathematically defined signal properties. Existing research that uses objective sound features extracted from audio signals include: the peak level [Perez-Gonzales and Reiss, 2008], differences in the spectra [Perez-Gonzales and Reiss, 2007], the spectral histogram of a mix [Kolasinski, 2008], and the full time-domain mix [Barchiesi and Reiss, 2009, 2010].

Research that uses objective sound features extracted from acoustic signals include: the measured level of background noise [Dugan, 1975, 1989], the difference in level between front and rear facing capsules [Julstrom and Tichy, 1976], and the relative sound levels of the components in a mix [Terrell and Reiss, 2009].

Music information retrieval (MIR) is a field of research that has grown rapidly in the past decade, since the first International Society for MIR conference in 2000. MIR, as its name suggests, concerns the analysis of music signals with the goal of extracting information from the content, and as a result, many musical features have emerged from it. The output of MIR work, i.e. the features it provides, are typically perceptual sound features, but there are many intermediate steps in their modelling that use low-level objective audio features, which, thanks to the MIR community, have been defined and classified. Lartillot et al. [2008] produced the MIR Toolbox, which as well as defining and describing features, provides open source Matlab code to calculate them.

### 1.2.2 Perceptual sound features

Perceptual sound features are evaluated using models of human perception, and are extracted from both acoustic and audio signals. Many facets of sound perception are dependent upon the listening level, which is not available if the related features are extracted from audio signals. Models that operate on audio signals either make assumptions with regard to the listening level, or they ignore its effect altogether. The most sophisticated models that operate on acoustic signals incorporate individual aspects of the human auditory system and are discussed in Section 1.3. The remainder of this section focusses on perceptual sound features extracted from audio signals.

Loudness is the perceived intensity of a sound, and is arguably the most well-studied perceptual sound feature. A recent loudness feature adopted by broadcasters is the European Broadcast Union (EBU) loudness unit, dBLU [Union, 2011]. Their model uses a frequency weighted filter that cuts the low frequencies and boosts the high frequencies, which based on their subjective evaluations, approximates the frequency sensitivity of the human auditory system[6]. The primary feature output by the model is the mean loudness of an audio signal, which is calculated after gating to remove low level parts of the signal, and reflects the tendency of a listener to focus on the louder parts of a sound when making overall loudness judgements. Similar loudness features were used by Perez-Gonzales and Reiss [2009a,b] for automatic mixing applications, and more

[6]The high frequency boost is not consistent with auditory theory, which is discussed in Section 1.6.

recently by Mansbridge et al. [2012].

The tools produced by the MIR community are designed to enhance and enrich our interaction with, and use of, music, and a widely used example of a tool is one that automatically generates song playlists. The MIR researcher does not know the listening conditions of the end user[7], so most, if not all, perceptually orientated MIR features operate on audio signals. Probably the most well known MIR features are Mel-frequency cepstrum coefficients (MFCCs) [Davis and Mermelstein, 1980], which are used to describe musical timbre. Mel-frequency refers to a frequency warped scaling, known as the Mel-scale, which spaces frequency points based on equal pitch distances, where pitch is the perceived frequency of a sound. MFCCs are a core feature in content-based MIR systems, for example: music recommendation [Logan, 2000], playlist generation [Pampalk, 2006], genre classification [Tzanetakis and Cook, 2002] and instrument recognition [Eronen and Klapuri, 2000]. There are a myriad of other MIR features, and the reader is referred to the MIR Toolbox [Lartillot et al., 2008] for a comprehensive list.

## 1.3 The human auditory system

In this section a brief overview of the human auditory system is provided, followed by specific references to the literature where the separate parts have been examined in detail. The reader is referred to [Moore, 1997, Pickles, 2008], which are comprehensive texts on the human auditory system.

The peripheral auditory system, i.e. the part of the auditory system prior to cognitive brain functions, is split into three distinct sections: the outer, middle, and inner ear. The outer ear covers the region from the external pinna to the ear drum, which acts as a direction dependent linear filter, and a resonator, reinforcing frequencies in the mid-range of our hearing. The middle ear contains three small bones, collectively known as the ossicles, which reside in the tympanic cavity. They form a hammer, anvil and stirrup like structure that transfers the vibrations of the ear drum into a piston like force that acts on the oval window, which is the boundary between the middle and inner ear (cochlea). The middle ear acts as a linear filter[8], and has its highest response in the mid-range.

The inner ear, known as the cochlea, is a fluid filled, spiral structure that converts the pressure

[7] A recently proposed audio format may make this possible [Terrell et al., 2012].

[8] At high sound levels approaching the threshold of pain the middle ear reflex applies attenuation, but the linear assumption is applicable to sound levels below this threshold

fluctuations induced in the oval window into nerve impulses that are sent to the brain. The cochlea contains the basilar membrane, which is lined with inner and outer hair cells. Action of the ossicles on the oval window causes pressure fluctuations to travel along the basilar membrane that are detected by the inner hair cells, causing them to fire. Pressure fluctuations with larger amplitudes produce a higher rate of firing, which determines the perceived intensity of the sound (loudness), and the relationship between firing and perceived intensity is known as the rate-level function. If the pressure fluctuations are very low, the induced rate of firing falls below the inherent random firing of the hair cells, and the sound is not detected. The minimum rate of firing for detection defines our threshold of hearing. In addition, there is a maximum rate of firing known as the saturation point. The cochlea acts as an amplifier and a compressor, and the compression gives us a logarithmic perception of loudness with respect to sound intensity.

There are two distinct yet complementary approaches that have been used to study the auditory system. The first applies psychophysical experimental methods [Stevens, 1975] to determine the relationship between stimuli and sensation. The second uses direct physiological measurement by probing different areas of the auditory system, and due to its intrusive nature is mostly restricted to studies of non-human mammalian subjects, or human cadavers. A third, relatively recent approach to studying the auditory system resides in the field of neuroscience, and uses brain imaging techniques such as functional magnetic resonance imagine (fMRI) [Alain et al., 2001] and Electroencephalography (EEG) [Huotilainen et al., 1998]. Work in this area is not included in this thesis. Early studies of the auditory system relied solely on psychophysical methods, so in order to track the chronology of auditory research, a brief overview of the main methods follows.

## 1.4 Psychophysical methods

Psychophysics is a field within experimental psychology and is the study of the relationship between stimulus and sensation. Psychophysical methods have evolved over the years, and three have reached prominence: the method of adjustment, magnitude estimation and the just-noticeable difference for discrimination or detection.

### 1.4.1 Method of adjustments

An experiment using the method of adjustment presents the participant with two stimuli. The participant is required to adjust an objective feature of one stimuli (the target), such that it has the same perceptual quantity as the other (the reference). If the perceptual quantity of interest was the loudness of a sound, the participant would apply a gain to the target to modify its objective intensity, such that it was equally loud when compared to the reference.

### 1.4.2 Magnitude estimation

In a magnitude estimation experiment the participant is presented with one or more stimuli and must estimate the magnitude of the perceptual quantity of interest. There are a number of variants in the type of magnitude that the participant must estimate (see Table 2, Stevens [1975] for more details). These are: nominal, ordinal, interval and ratio estimation. With *nominal* estimation the participant must place each stimulus into a classification system with a predefined structure. This has been used extensively when measuring linguistic acceptability, where participants make judgements on the grammatical correctness of sentences, i.e. unacceptable, acceptable, good, excellent etc (see Bard et al. [1996] for a discussion of all forms of magnitude estimation in linguistic research). With *ordinal* estimation, the participants rate the magnitudes of the sensations, and place them on an ordinal scale. This technique was used many hundreds of years ago to classify the brightness of stars. With *interval* estimation, the participant estimates the magnitudes of sensations and places them on a continuous scale, e.g. the temperature scales, Celsius and Fahrenheit. Finally, with *ratio* estimation the participant is presented with two stimuli and must assign numerical values to describe the relative magnitudes of their associated sensations, e.g. the participant hears two sounds and rates the first as being twice as loud as the second, and assigns values of 10 and 5 respectively (the participant is free to use any number system they are comfortable with). The benefits of interval and ratio estimation is that they allow mathematical operations to be performed on the test results. With *interval* estimation, addition (and subtraction) can be performed, e.g. if sensation A is 10 units larger than B, and sensation C is 10 units larger than A, then we know that C is 20 units larger than B. With *ratio* estimation, multiplication (and division) can be performed, e.g. if A is twice as large as B, and C is twice as large as A, then C is four times as large as B. If an *interval* scale is absolute, e.g. as with the Kelvin scale for temperature, then multiplication can also be performed.

### 1.4.3 Just noticeable difference

The just-noticeable difference (JND) is the minimum change in a given stimulus property that can be detected by a participant. There are two types of JND test, discrimination and detection. In the former, the participant must discriminate between two stimuli to, for example, identify which of two sounds has the higher intensity. In the latter, the participant must detect some property of the stimuli, and its application includes determining the threshold of hearing, in which one stimuli is silence, and the other is a sound. The psychometric function shows the relationship between a given change in the stimulus and the probability that it will be detected.

JND tests have been updated to incorporate an adaptive procedure, that hones in on a specific point on the psychometric function. Within such tests, multiple evaluations are made, and for each, a correct response causes the difference between the two stimuli to be reduced. For the discrimination task described above, this would involve reducing the difference in intensity between the two sounds. An incorrect response is followed by an increase in the difference between the stimuli. This process is known as an adaptive, forced choice, up-down procedure, and for a given experiment is repeated until some convergence criteria are met, upon which time the average of all direction changes is used as an estimate of the JND. The point on the psychometric function that this corresponds to depends upon the probability of giving a random correct answer, and is 50% for the process described above. The up-down method was adapted by Levitt [1971], so that a down-step is only taken if three correct responses are performed sequentially, which moves the JND value to the 71% point on the psychometric function. This is still best practice today.

## 1.5 Early work on auditory models

Research into perceptual sound features from the first half of the 20th century is still commonly cited today. This work was driven entirely through experimentation using psychophysical methods, which at the time were not as widely accepted as they are today, so it is worth covering this work separately, as a tribute to those early pioneers.

Fletcher and Munson [1933] used the method of adjustment to provide the definition of loudness level, which is the gain applied to tones of different frequencies, such that they are equally loud when compared to a 1kHz tone of a given intensity. The most well known output from this work was the equal loudness contours. Fletcher and Munson [1933] also suggested a relationship between loudness and the rate of nerve firing, which has since been supported using

physiological measurements. Fletcher and Munson [1937] examined the relationship between loudness and masking, and proposed the excitation pattern model of sound perception, which describes the mechanism by which sounds of different frequencies excite different parts of the basilar membrane. Fletcher [1940] modelled the basilar membrane as a series of auditory filters. He showed that if a signal was too far from the centre frequency of an auditory filter, then it was rejected, and this lead to the interval within which sounds were accepted to be defined as the critical bandwidth. Zwicker et al. [1957] verified the excitation pattern and critical bandwidth models using the method of adjustment for complex tones[9]. Zwicker showed that there was a sudden increase in the perceived loudness when the separation of the tones reached a critical point, i.e. the edge of the critical band, and Greenwood [1961] demonstrated that the widths and frequencies of the critical bands were directly related to their position on the basilar membrane. Since then, many researchers have sought an equation to describe the relationship between filter frequency and bandwidth, which is described as the equivalent rectangular bandwidth (ERB). At a similar time, Stevens [1957] derived a psychophysical power law relating stimulus intensity to perceived loudness, which was related to the compressive action of the cochlea.

## 1.6 Auditory theory

In this section, the literature is discussed with respect to the three parts of the auditory system: the outer, middle, and inner ear.

### 1.6.1 The outer ear

The sound pressure at the ear drum is expressed as a transfer function, relative to the sound in the free-field. In general, the free-field sound is measured using a microphone positioned where the centre of the listener's head would be. Shaw [1974] collated the measurements from prior studies, and produced best-fit transfer functions that modelled the outer-ear as a function of frequency and azimuth angle (the direction in the horizontal plane). Differences in response at different azimuth angles are due to the pinna, and are most prominent above 6kHz, as demonstrated by Kuhn [1979] using manikins. The transfer function for a sound at frontal incidence (taken from Shaw [1974]), is shown in Figure 1.2. It shows substantial gain in the mid-range, with a peak at around 1 to 4 kHz, which coincides with the range of human speech.

[9] A complex tone is a tone made up from two or more sinusoids of different frequencies.

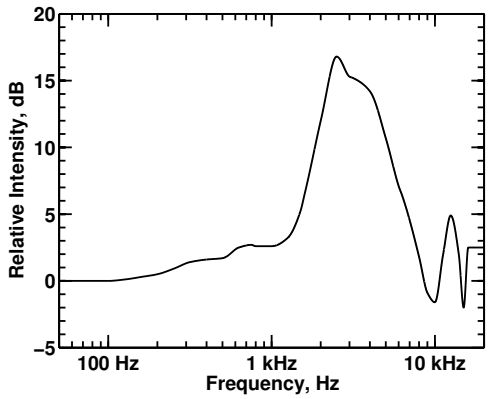


Figure 1.2: The outer-ear transfer function converting the free-field sound intensity to the intensity at the eardrum, as given by Shaw [1974].

### 1.6.2 The middle ear

The middle ear is enclosed, so it is not possible to use probes to measure sound pressure or velocity directly. Measurements of the middle ear transfer function therefore come from human-cadaver ears, and, due to the mechanical nature of the ossicles, it is argued that its operation does not change much after death. Puria et al. [1997] measured the magnitude response characteristics of four human-cadaver ears in the frequency range from 50 Hz to 12 kHz, and found a peak gain of 20 dB from 500 Hz to 2 kHz, relative to the response at 50 Hz. Below this range the slope was 4 dB/octave, and above it was -8 dB/octave. Aibara et al. [2001] measured sound pressure and velocity in human-cadaver ears, from which they were able to calculate the impedance of the middle-ear (complex transfer function). They found a peak magnitude response of 23.5 dB at 1.2 kHz, relative to the response at 50 Hz, with a 6 dB/octave slope below and -7 dB/octave slope above, and the phase angle was shown to have a slope of -87$^o$/octave. The middle-ear transfer function implemented in the loudness model of Moore et al. [1997] (to be covered in detail later) made minor adjustments to the data from Puria et al. [1997][10], and has since been implemented as an ANSI standard [**?**], and is plotted in Figure 1.3.

### 1.6.3 The inner ear

Fletcher [1940] was the first to model the basilar membrane as an auditory filter bank. Since then, much work has gone into determining auditory filter shape and spacing. Patterson [1974, 1976] used an adaptive JND method to measure detection thresholds of tones in the presence

[10]Their measurements were made with the ear canal open, and the adjustments by Moore *et al.* sought to correct the errors in the 2-3 kHz regions that this introduced.

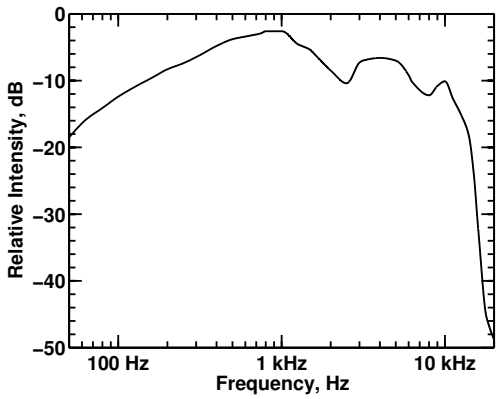


Figure 1.3: The middle-ear transfer function, which converts intensity at the ear drum to intensity at the oval window, as given by ANSI S3.4-2007 **?**.

of masking noise. In Patterson [1974], tones were placed at different frequencies within a narrow band noise. As the tone moved further from the centre frequency of the noise the detection threshold decreased, because the tone was beginning to excite the neighbouring auditory filter. In [Patterson, 1976], the width of a notch in a noise band centred about a tone was varied, which showed that the measured thresholds were lower with wider notches. The shape of the auditory filter was estimated from the derivative of the function relating threshold to: distance from noise centre frequency [Patterson, 1974], and notch width [Patterson, 1976]. He showed that a Gaussian curve gave a good approximation to the filter shape within the critical bandwidth, and was in agreement with the earlier work of Zwicker et al. [1957]. The filter shapes at this time were assumed to be symmetrical, and it was argued that this was a reasonable assumption for the low sound intensity used in the tests.

The effect of level on auditory filter shapes was examined by Weber [1977]. He showed that for noise with narrow band notches, level did not change the detection threshold, but that for noise with wide notches the threshold was lower at high levels. The lower thresholds represent an increase in the signal to noise ratio, and this was interpreted as an increase in the auditory filter width. Patterson and Nimmo-Smith [1980] investigated asymmetry in the audio filters by placing the notch off-centre with respect to the tone. This revealed differences in the detection threshold when the notch was centred above and below the tone, and the auditory filter shapes were subsequently modelled using a pair of exponentials, either side of the tone frequency, with different decay rates, i.e. the auditory filters were asymmetric. In a similar study, Patterson et al. [1982] defined the auditory filter shapes as rounded exponentials, referred to as the ‘roex’ filter

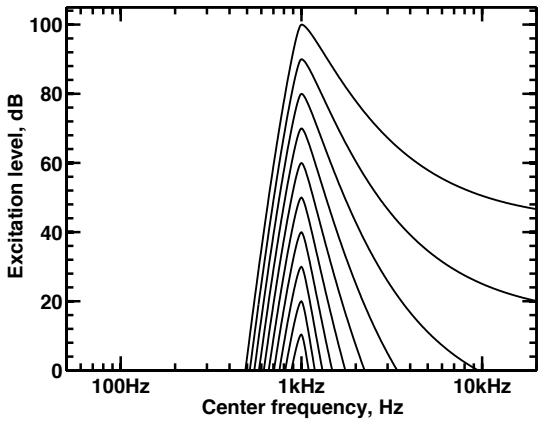


Figure 1.4: Illustration of the 'roex' filter shapes for excitation levels between 10 and 100 dB, in 10 dB intervals. The excitation level is expressed relative to the reference excitation caused by a 1 kHz sinusoid at 0 dBSPL.

shape. This was verified in a similar study by Glasberg et al. [1984], but was performed over a larger dynamic range; and Moore and Glasberg [1987] provided equations that could be used to calculate excitation patterns of sounds as a function of their frequency and level. Glasberg and Moore [1990] performed a comprehensive study of level dependency and asymmetry in auditory filter shapes, and incorporated the effect of the headphone response, and transfer functions of the outer and middle ear. Using the output of this study, the excitation pattern from a 1 kHz pure tone at different intensities is shown in Figure 1.4. The filter shape is asymmetrical, and higher intensities cause the excitation above threshold to extend to higher frequencies. This effect is known as the upward spread of excitation.

The cochlea acts as a compressor, which gives us a logarithmic perception of loudness with respect to sound intensity. This was demonstrated by Stevens [1957] by extracting the psychophysical power law relating intensity to loudness. The power law is a fundamental psychophysical law [Stevens, 1975], and states that when a stimulus intensity is doubled, the sensation also doubles. Magnitude ratio estimation is used to determine these relationships, and Stevens showed that sound intensity (E) and loudness (N) were related by the equation[11],

$$N = E^{0.3}. \tag{1.1}$$

The mechanism of cochlea compression is attributed to the motive action of the outer hair-cells, in particular their synchronous interaction with the inner hair-cells. This mechanism is

[11]The exponent of $E$ is less than 1, so the function is compressive.

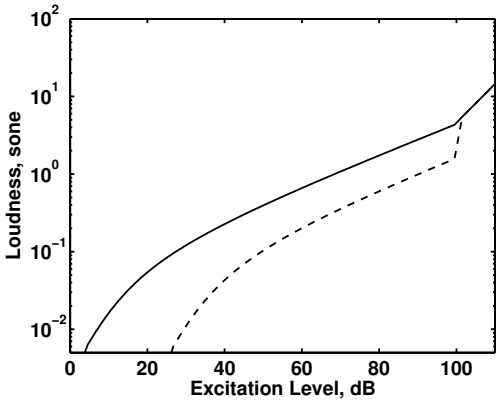


Figure 1.5: The compressive function of the cochlear that converts sound intensity into loudness (per frequency band). The solid line is for frequencies of 500 Hz and above, and the dashed line is for frequencies of 40 Hz. Compression curves at intermediate frequencies lie between these two curves.

active, in that it is a form of amplification that is driven from within the auditory system, and was first reported by Kemp [1978]. Knowledge on this mechanism has been complemented by studies of non-human, mammalian physiology. Early work by Rhode [1971] produced cochlear input-output functions measured for live squirrel monkeys, and similar measurements were taken from chinchillas by Robles et al. [1986]. The specific action of the outer hair cells has been observed as a contraction of their length, which causes changes in their mass distribution and stiffness properties. Neely [1993] developed a model of the process, by variably tuning the admittance of outer hair cell coupling with inner hair cells, along the length of the basilar membrane. Cochlea amplification was further verified using psychophysical methods by Oxenham and Plack [1997], who showed that there was most compression for sound levels between 50 and 80 dBSPL. They extracted a power law exponent of 0.16 that was shown to be in agreement with physiological studies. In the later chapters of this thesis, the loudness model of Moore et al. [1997] is implemented, and the cochlear compressor function they used is illustrated in Figure 1.5. Their model of cochlear compression includes the absolute thresholds at different frequencies, which are constant above 500 Hz, but which increase below it. The solid line in Figure 1.5 is the compression curve for frequencies above 500 Hz, and the dashed line is for 40 Hz.

### 1.6.4 Loudness models

Although other perceptual quantities exist, for example pitch and timbre, loudness is arguably the best understood. Loudness is the perceived intensity of a sound as measured in sone. The

sone scale is defined such that a 1 kHz tone at 40 dBSPL has a loudness of 1 sone. Prior automatic mixing work has often used either pseudo-perceptual loudness features, or objective level features. They are clearly both related to loudness, and so extending the automatic mixing work to use a fully perceptual loudness feature is a logical next step. Following is a discussion of loudness models, from which loudness features can be evaluated.

A loudness model takes as input a sound (an acoustic signal scaled to an absolute pressure), and outputs loudness. The early work from Fletcher and Munson [1933] produced a model of equal loudness for tones, which output the loudness level of a tone as a function of its frequency and intensity. Their excitation pattern model has since been implemented using an auditory filter bank, and compressive cochlea input-output functions, which allows the loudness of a steady-sate sound to be estimated [Moore et al., 1997, Zwicker and Fastl, 1990, Zwicker and Scharf, 1965]. The functionality of these models is broadly similar. The sound is first passed through linear filters to account for the outer and middle ear responses, and is then split into frequency components corresponding to the auditory filter centre frequencies and bandwidths. The excitation pattern is evaluated for the sound components in each band, and are then summed to give the total excitation pattern. The excitation pattern is converted into specific loudness, which is the loudness per auditory filter band, before being integrated across all frequencies to give the overall loudness of the sound. The model from Moore et al. [1997] gave better agreement with experimental data compared to the earlier models, and also provided a means to estimate the partial loudness of a sound, which is its loudness when heard in the presence of other masking sounds. The partial loudness of a sound is a very important concept when describing a mix, because the component sounds are heard simultaneously and can therefore mask one another.

Differences in the perception of loudness for steady-state and time-varying sounds were first identified in experiments that sought to match the loudness of two-tone complexes. Zwicker and Fastl [1990] found that closely spaced tones within a complex would interact and cause amplitude modulation, known as ‘beating’. The loudness of a steady-state sound was known to be related to its RMS intensity, but this was not found to be the case for modulated sounds, which had a lower RMS intensity compared to their steady-state counterparts when equally loud. This suggested that the loudness of a modulated sound may be related to its peak intensity. Similar effects were reported by Bauch [1956]. However, work by Moore et al. [1998] showed the opposite effect, i.e. that modulated tones required a higher RMS intensity to be equally loud when compared to a

steady-state tone. Zhang and Zeng [1997] attempted to replicate the results of Zwicker and Fastl [1990] using an adaptive JND method to remove the effect of bias in loudness matching tasks. They showed the same direction of change as Zwicker and Fastl [1990], but the effects were far less pronounced. Hellman [1985] carried out similar tests using magnitude estimation on two-tone complexes combined with a low-pass noise. Hellman's data agreed with Zwicker's results, showing that for small frequency differences between two tones, the beating-induced modulation caused an increase in the loudness for equivalent RMS intensity.

Moore et al. [1998] sought to clarify the effect of amplitude modulation on the loudness of both tones and noise, using an adaptive JND discrimination method. They identified areas in previous studies that may have introduced biases, for example, the experiments that used the method of adjustment procedure only required the participant to adjust the level of the modulated sound. Performing the adjustment introduces biases, so experiments that use this method must be repeated with both types of stimuli (i.e. steady-state and modulated sounds) being subject to the level adjustment. A similar issue was highlighted with the adaptive JND method used by Zhang and Zeng [1997], who only applied level adjustment to the steady-state sound. To overcome these issues, Moore et al. [1998] used an adaptive JND method, with level adjustment on both the steady-state and modulated sounds. They show clear bias effects depending on which sound was level adjusted, and this bias was also shown to be dependent on the listening level. The bias was removed by taking the mean of the two results. Statistically insignificant differences were found between the loudness of modulated and steady-state tones with equal RMS intensity. However, they suggest that for large modulation depths, the compressive input-output function of the basilar membrane may account for differences in the loudness of steady-state and modulated sounds, and that the loudness is related to the RMS intensity after compression. This was validated by repeating the experiment on hearing impaired listeners, whose impairment was assumed to be an absence of basilar membrane compression.

The studies above demonstrate the difficulties in estimating the loudness of time-varying sounds. Although the effects of amplitude modulation were clarified by Moore et al. [1998], the conclusions are not much use in estimating the loudness of musical sounds. Subsequent models of loudness for time-varying sounds apply a steady-state loudness model on a frame by frame basis, to provide a time-varying loudness function, i.e. the input is a pressure time-function, $p(t)$, and the output is a loudness time-function, $L(t)$. Zwicker [1977] suggested this approach

and provided the procedure to build a loudness measuring device based on his model of steady-state sounds [Zwicker and Scharf, 1965]. For the measurement of speech, the system output was compared to subjective data on the loudness of speech-like noise, and Zwicker concluded that the overall loudness impression of speech is related to the peak of the loudness time-function. This is an interesting conclusion, because it provides a loudness feature (the peak loudness) that can be extracted from the loudness time-function, to give an impression of the overall loudness. A similar modelling approach was taken by Glasberg and Moore [2002], who implemented an updated version of their steady-state loudness model [Moore et al., 1997], on a frame by frame basis. Specific loudness and loudness for each frame were termed instantaneous specific loudness, and instantaneous loudness respectively, which are intermediate loudness quantities present in the model, but which are assumed to occur too quickly to be perceived. Glasberg and Moore [2002] used temporal integration to convert the intermediate, instantaneous quantities into the perceptual measures of short-term loudness (STL) and long-term loudness (LTL), which incorporates the accumulation of loudness as discussed by Buus et al. [1997], Florentine et al. [1996]. The STL can be used to estimate the momentary loudness of a sound, and the mean LTL can be used to estimate the overall loudness impression of a sound. The mean LTL, like Zwicker's peak loudness, is another candidate loudness feature. Glasberg and Moore [2005] extended their model further to allow the thresholds of time-varying sounds to be predicted in the presence of background noise, which also included a model of partial time-varying loudness.

## 1.7 Automatic-mixing sound-features

A number of conclusions can be drawn about the sound features used in existing automatic mixing applications. Firstly, the excitation pattern is a function of level, and the conversion of intensity to loudness in the cochlea is non-linear, so our perception of sound intensity will be dependent upon the listening level. Second, when listening to a mix the component sounds are heard simultaneously, so the concept of partial loudness and masking become important. Neither of these phenomena can be accounted for when using objective levels, or pseudo-perceptual loudness features. Thirdly, there is no validated model that describes the overall loudness impression of a musical sound, and so although it is possible to express loudness as a time-function, there is no means to convert this into a single loudness quantity.

## 1.8 Thesis objectives

As with all automatic mixing research, the overriding objective of this work is to make the mixing process more accessible to amateurs, and easier to do well for professionals. It is the intention to base the work on perceptual sound features that are extracted use full auditory models, which operate on acoustic signals. Mixing is invariably more complex for live, compared to recorded music, due the presence of live acoustic sources. Whilst the aim is to unify approaches to both types of mixing, the issues that arise in live must be understood, and dealt with separately, and the objective is to reduce to complexity of live mixing to the same level as recorded mixing. Once this has been done, further developments can be applied to both.

The use of sound features extracted from audio signals has been cited as a potential weakness in existing automatic mixing work. This view is supported by auditory theory, due to the non-linear, and frequency dependent relationships between loudness and listening level (their effect on loudness perception of musical sounds is as yet unknown, and must be quantified). Furthermore, if the effects are proven to be significant, there are currently no validated models or features to describe loudness relationships between musical sounds, so these must be researched, and made available for automatic mixing systems.

The field of automatic mixing is relatively immature, and as a result, a number of different approaches have been taken (see Section 1.1). Although presented here under the umbrella of automatic mixing, the approaches discussed are distinct. Based on the outcome of the planned work, the current automatic mixing paradigm will be examined, with the aim of unifying all approaches, to provide a flexible framework for future work within this field. This will likely include a rebranding of 'automatic' mixing to a more generic term.

The objectives are enumerated below for clarity.

1. To develop a model of the mixing process, applicable to both live and recorded music, that can be used to describe the mix experienced by a listener as a set of acoustic signals.

2. To develop robust models and algorithms that can be used to do live automatic mixing in all conditions, using objective sound features.

3. To explore the effect that listening conditions have on the perception of loudness for musical sounds, and to evaluate the implications that this has on the use of objective and pseudo-perceptual sound features for describing a mix.

4. To provide a validated perceptual feature that describes the loudness relationships between the musical sounds within a mix, which are suitable for use in automatic mixing algorithms.

5. To incorporate the perceptual loudness feature into automatic mixing systems.

6. To examine the current automatic mixing paradigm, and provide a framework upon which future work can be based.

## 1.9 Thesis outline

In Chapter 2, two models are developed. The first model is a general model of mixing, which enables mixes to be described as a set of acoustic signals. It includes simple environmental-acoustic models that take into account the interactions between different sources and room acoustic effects. The second model is focussed on live performance, and models the act of mixing as a constrained multi-objective optimisation problem, in which the parameters are the gain controls on the mixing desk, the objectives relate to features of reference mixes that are to be recreated, and the constraint prevents the onset of acoustic feedback. Chapter 3 contains a case study that implements these models to do live, offline automatic mixing, and provides a robust two-stage algorithm to solve the automatic mixing optimisation problem. In addition, the limiting factors in providing the reference mixes are discussed, namely: the interdependency between mixes at different locations, the contribution of the live acoustic sources (i.e. the instruments) to the mix, and the feedback constraint.

In Chapter 4, the automatic mixing algorithm is used in venues of different sizes, to determine the effect of venue size on the limiting factors mentioned above. In general, the limiting factors are more restrictive in small venues, and conclusions are drawn on the absolute sound levels that can be accommodated, particularly in relation to drums. In large venues, the limiting factors have less effect (in particular, the contribution of sound from the live acoustic sources becomes negligible) and it is easier to provide the reference mixes. However, in large venues, more sound energy is required, which necessitates the use of multiple loudspeakers, grouped into arrays. Loudspeaker array optimisation is examined in Chapter 5, and a robust computational optimisation strategy is developed that minimises the differences in the frequency response at all audience locations. Loudspeaker optimisation is a necessary precursor to the production of live music since it provides a solid platform upon which subsequent automatic mixing is based.

The work up to and including Chapter 5 provides the means to control objective sound features for live automatic mixing and sound system optimisation. The effect of this is to unify work on live and recorded mixing, because the additional practical considerations of live music can now be accommodated automatically. Loudness features for musical sounds and mixes are examined experimentally in Chapter 6 using the psychophysical method of loudness ratio estimation. It shows that participants are able to reliably estimate the relative loudness of the component sounds in a mix, suggesting that relative loudness is a good feature with which to describe the perception of a mix, and related mixing tasks. Within this study, the loudness ratios are found to be strongly dependent on the listening level, creating doubt as to the suitability of objective features, or perceptual features evaluated from audio signals, for describing a mix. In Chapter 7, an existing loudness model for time-varying sounds is extended, incorporating a signal specific bias that describes the perceived loudness of musical sounds that is dependent upon their dynamic characteristics. The model is able to predict the experimental loudness ratios from Chapter 6, and is a substantial improvement on the state of the art.

In Chapter 8 the loudness model for musical sounds is used to define a new perceptual feature termed the *loudness balance*, which describes the loudness relationships between the sounds in a mix. The feature is applied to the live performance case study, demonstrating the limitations in objective features. A new perceptual mixing tool is outlined, suitable for both live and recorded music, that enables perceptual features to be controlled directly, inline with the re-parameterised audio effects discussed in Section 1.1.3. Furthermore, by applying the perceptual mixer as an analytical tool, a method is presented to extract statistical models of best practice, that can replace existing heuristic models, to do fully-automatic mixing. A final application is discussed, which is an audio transmission format that allows perceptual sound features to be recreated irrespective of the listening conditions. Conclusions, and a brief overview of future work are given in Chapter 9.

## 1.10 Summary

In this chapter the motivations behind the work in this thesis have been outlined, i.e. to make it easier to mix music. The field of automatic mixing, within which this work resides, has been discussed. A common theme has emerged, which is the need to describe the objectives of the mixing process using sound features. Most existing work uses objective or pseudo-perceptual

sound features that are extracted from audio signals. Additional perceptual features, which are evaluated using auditory models that operate on acoustic signals, have been detailed, with the intention of employing them at later stages in the research. The objectives of this thesis have been listed, and the contents have been outlined.

# Chapter 2

# The Music Production Model

In the previous chapter, existing automatic mixing research was discussed. Almost all existing work describes the mix using sound features extracted from *audio* as opposed to *acoustic* signals, and the limitations of this approach at incorporating mix perception have been discussed. In this chapter a general model of mixing is proposed, with an emphasis placed on live performance, where the model is inherently more complex because of the presence of live acoustic sources. A set of assumptions are stated, which form a framework upon which automatic mixing can be based. The model and framework is then developed further, to specifically apply to offline, live mixing; and includes models of source radiation and receiver response patterns, and room acoustic effects.

## 2.1 General music production model

In this section the general model of mixing is outlined, which is applicable to both live and recorded music. The acoustic and audio signal paths are modelled using the flow chart shown in Figure 2.1. The music is performed by a group of performers each of whom plays an instrument (the voice is considered an instrument). Each instrument produces a sound in the form of an acoustic signal and/or produces an audio signal (a synthesiser produces only an audio signal). In some instances an instrument, for example a drum kit, will output multiple sounds, and if this is the case it is treated as multiple instruments. Receivers are either human listeners, i.e members of the audience or performers, or microphones at a live event[1], or the mixing engineer for recorded

[1]The signal received by the microphone is important when considering acoustic feedback.

music.

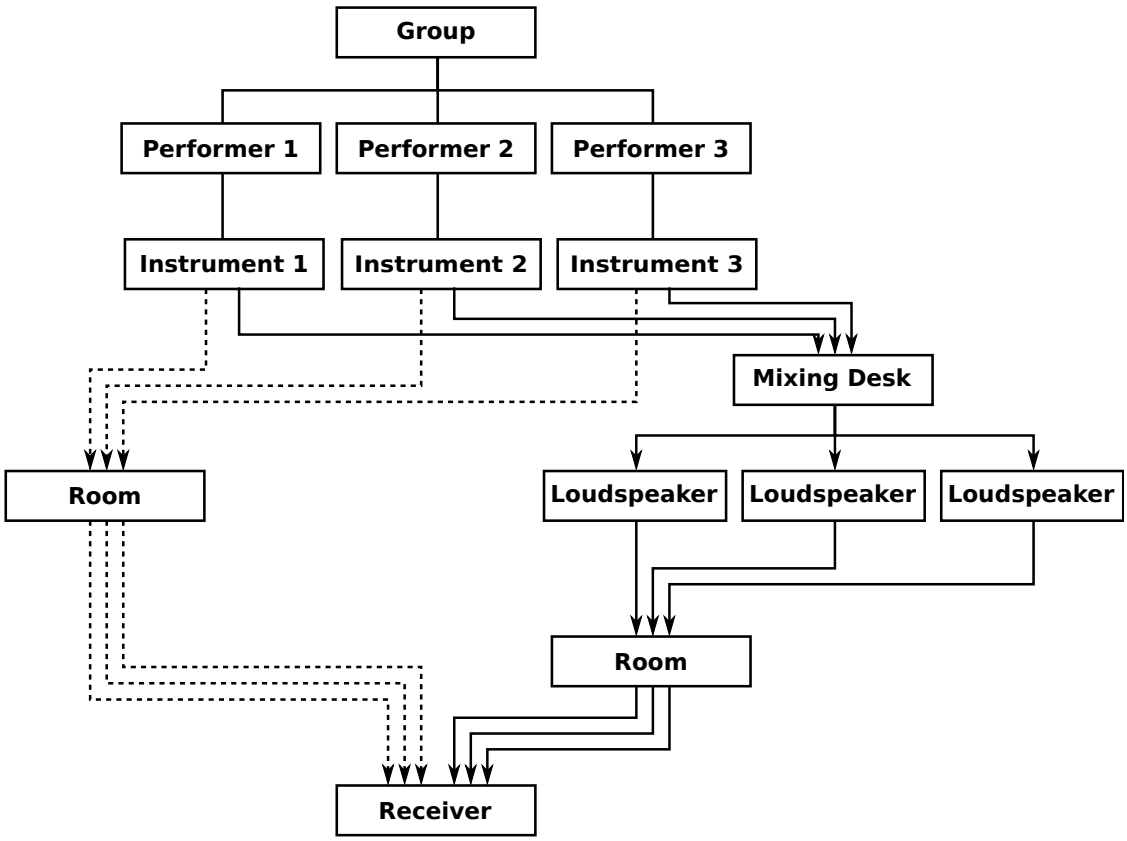


Figure 2.1: The general model of mixing. The group is composed of a number of performers each of whom plays one or more instruments. The dashed arrows represent the direct signal path from instrument to receiver and the solid arrows represent the reinforced signal path from instrument to receiver, via the mixing console (where signal processing is applied to the audio signals), and sound reinforcement system.

There is a direct acoustic signal path from the instrument to the receiver (indicated by the dashed lines and arrows on the left side of Figure 2.1) and a reinforced signal path via the mixing desk and sound reinforcement system (indicated by the solid lines and arrows on the right side of Figure 2.1), the outputs of which are loudspeakers. The reinforced signal path includes both acoustic and audio signals. The acoustic signal path from an instrument or a loudspeaker to the receiver are affected by the acoustic medium and the room. For recorded music there is typically no direct signal path.

The acoustic signals are input to the mixing console as audio signals, via an acoustic signal path and a pressure transducer (i.e. a microphone), or an electrical direct input (DI). In the case of recorded music, it is likely that the mixing stage is done separately from the recording stage, i.e. the audio signals are recorded and stored within the mixing system. Once inside the mixing desk, the audio signals can be transformed using signal processing tools, for example

gain and equalisation, before being output by the loudspeakers. The acoustic signal that reaches the receiver is a linear superposition of all signal paths from all instruments. It is a mixture of musical signals, and is the definition of a mix used in this thesis.

The proposed model holds equally well for the reproduction of recorded music, i.e. via iPod, nightclub, home stereo etc, as well as the mixing process. In these instances, the mix is stored as an audio signal, e.g. on a CD, and is passed through an audio processing device, e.g. a home stereo, and is output by loudspeakers as an acoustic signal, with the signal reaching the listener being subject to room acoustic effects.

## 2.2 Live musical performance

The model presented in the previous section is applicable to any form of music production or reproduction, but is most complex when applied to live music. The main reasons for this are the presence of live acoustic sources, and the need to provide mixes to many different receivers, i.e. to the audience and the performers, who are positioned at different locations within the venue. A live performance therefore is the 'worst case scenario' when attempting to model the mixing process, hence if the problems associated with live music can be tackled, it will be relatively simple to apply the same principles to non-live applications. Therefore, for the remainder of this chapter, and the subsequent three chapters, the focus is on live musical performance.

### 2.2.1 Front of house and monitor mixes

The mix experienced by the audience and the performers are called the front of house (FOH) and monitor mixes respectively. A uniform FOH mix is desirable for the audience locations, whereas each performer has a different ideal monitor mix. The sound reinforcement system contains designated FOH and monitor loudspeakers. In large venus, the distances between the instruments and the audience causes the direct signal path to have minimal contribution to the FOH mix. In smaller venues it can be very significant. The close proximity of the performers to their instruments means that the direct signal path invariably contributes to the monitor mixes, even in large venues.

The mixes experienced by the listeners are interdependent, i.e. it is not possible to make changes to one mix without affecting all others. In the coming chapters this is referred to as coupling between mixes. In larger venues the increased distances between receivers, instruments

and loudspeakers means that coupling between FOH and monitor mixes is reduced. In smaller venus strong coupling can exist between all mixes. Mixes cannot be set independently and a best fit to the requirements of all listeners must be sought.

### 2.2.2 Controlling the mixes

It is the responsibility of the mixing engineer to provide mixes to each listener that fit the aesthetic requirements of the music. This is done by adjusting the control parameters on the mixing desk, such as gain and equalisation, to modify the audio signals in the reinforced signal path. These changes can only be made while experiencing the mix at the mixing desk location (it is possible for the engineer to check multiple locations but the controls cannot be adjusted while away from the desk unless a remote control is available). The engineer cannot experience the mixes at the performer locations and so can only infer the required changes in the mixes through communication with the performers.

The changes that can be made to the control parameters are constrained. Two constraints present at all musical performances relate to the maximum allowable sound level, and the onset of acoustic feedback. The former is more restrictive at venues in residential areas that have a lower sound level limit. The latter restricts the transformations that can be applied to the audio signals that have been input to the mixing desk via a microphone. It is problematic with instruments that produce an acoustic signal with a low sound level, and in popular music this is most common with vocals.

## 2.3 The engineer's role as a optimisation problem

The task with which the engineer is faced can be viewed as a constrained, multi-objective optimisation problem. Each objective relates to the mix at a given receiver location, but because the mixes are coupled they cannot be considered independently. The constraints relate to the maximum allowable sound level and acoustic feedback, as discussed in the previous section, and the parameters within the optimisation problem are the controls on the mixing desk. Three assumptions are made at this point:

1. The performance can be modelled accurately, i.e. if the instrument signals and mixing desk controls are known it is possible to evaluate the mix at any location.

2. The mix at a given location is *defined* as the linear superposition of multiple acoustic signals at that location, and a particular mix can be *described* using a set of features extracted from the acoustic signals from which it is composed. No assumption is made at this stage with regard to the form of these features.

3. The set of features which describes the desired mix at each listener location, hereby referred to as the reference mix, is available.

If these assumptions hold, then it is possible to formalise the optimisation task undertaken by the engineer. The degree to which each objective is satisfied can be evaluated by comparing the features of the actual mix, with the features of the reference mix at each listener location. This gives a numerical measure of the quality of the mixes. The total error is evaluated by combining the error measurements of all objectives. The set of parameters i.e. the controls on the mixing desk, can then be sought to minimise this error. This will replicate the role of the engineer, and will enable the mixing desk controls to be set automatically to deliver the features of multiple target mixes to multiple listener locations during a live performance. These three assumptions form a framework upon which any form of automatic mixing can be based.

## 2.4 Sources and receivers

Instruments and loudspeakers produce acoustic signals and are collectively referred to as acoustic sources. Instruments can also produce audio signals that are input directly to the mixing desk, but these are omitted from the model at this stage. The definition of an instrument includes sources of amplification that are specific to the instrument i.e. a guitar amplifier is part of the electric guitar instrument. All acoustic sources are modelled as point sources from which the acoustic signal is radiated, and is described using a radiation pattern, which is generally frequency dependent, and can be nonlinear with respect to the sound level.

Idealized, theoretical models of loudspeaker radiation patterns have been presented in the literature, for example the piston in an infinite baffle [Beranek, 1954]. For non-theoretical applications such as loudspeaker line array modelling, the radiation patterns of loudspeaker are measured [Meyer, 1984a]. The radiation pattern is represented by a set of impulse responses for a discrete set of radiation directions, measured at a reference distance beyond which the radiation pattern of the sound field can be treated as being uniform. Inside the reference distance, termed the near-field, the radiation pattern is not uniform due to interactions between different

parts of the acoustic source (i.e. between multiple transducers in a multi-way loudspeaker), and interactions between the acoustic signal and the geometry of the acoustic source (i.e interactions with the loudspeaker cabinet). Radiation patterns can be measured accurately for loudspeakers because they can be driven by a test signal, and the response can be measured. This is not the case with instruments such as a voice or a drum kit which do not have their own separate amplification system. Even if such radiation patterns were available for all instruments, difficulties would still arise because the microphones used to convert the acoustic signals into audio signals are generally placed within the instrument near-field, where the radiation pattern is non-uniform.

The radiation patterns of acoustic sources and loudspeakers are simplified to 2-dimensional polar broadband gain functions, i.e. the energy is a function of radiation direction, but not a function of frequency. If $x_I(0)$ is the on-axis acoustic time domain signal produced by an instrument at a distance of 1 m, then the acoustic signal at an off-axis angle of $\theta$ is given by,

$$x_I(\theta_I) = x_I(0)R_I(\theta_I), \tag{2.1}$$

where $R(\theta)$ is a polar broadband gain function, and the subscript $I$ identifies the sources as an instrument.

Figure 2.2 shows the signal path from an instrument to a loudspeaker. It shows the instrument, the microphone, the gain on the microphone input $g_M$, the mixing desk which applies a linear transfer function $P$, the loudspeaker amplifier gain $g_L$, and the loudspeaker. Instrument acoustic signals are sent to the mixing desk via a designated microphone. Bleed from other instruments is omitted from the model and microphones are identified by the instrument to which they are assigned (i.e. microphone 1 is assigned to instrument 1).

It is assumed that the system has been calibrated, by adjusting $g_M$ and $g_L$, such that the signal at the loudspeaker on-axis reference distance, is identical to the signal at the instrument on-axis reference distance, when $P$ is a unity broadband gain function. $g_M$ and $g_L$ are static controls external to the mixing desk and are not included as parameters in the automatic mixing algorithm. $x_{L0}$ and $x_{L\theta}$ are the acoustic signals at the loudspeaker on-axis and off-axis reference distances respectively. In the frequency domain where $X$ is the Fourier transform of $x$,

$$X_L(\theta) = X_L(0)R_L(\theta) = X_I(0)PR_L(\theta), \tag{2.2}$$

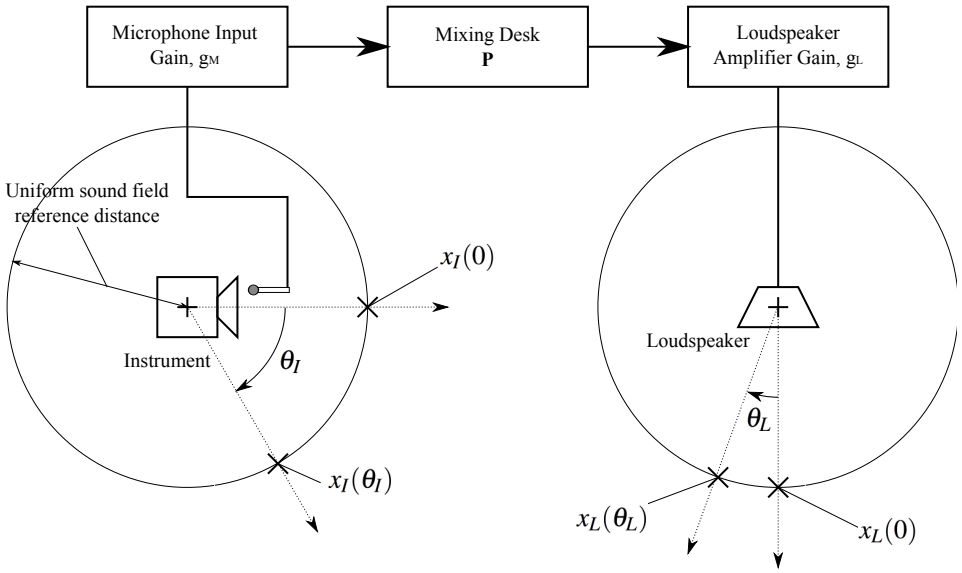


Figure 2.2: An illustration of the audio signal path from instrument to loudspeaker. It is assumed that the microphone gain and loudspeaker amplifier gain are set such that the on-axis signal 1m from the loudspeaker is equal to the on-axis signal 1m from the instrument when the mixing desk controls are set so that the signal entering the mixing desk is exactly equal to the signal leaving it.

where $R_L(\theta)$ is the polar broadband gain function describing the loudspeaker radiation pattern, and the subscript $L$ identifies the sources as a loudspaeker.

The responses of the receivers are also modelled using polar broadband gain functions, given by $A_R$, and in addition, the listener responses are modelled as monaural. The source radiation, and receiver response, polar broadband gain functions, are shown in Figure 2.3. The loudspeaker and amplifier radiation patterns are based on Meyer loudspeakers and are approximated using Meyer's Mapp online sofware [Meyersound, 2011]. Vocal and drum radiation patterns are modelled as omnidirectional along with the listener response, and the microphone response is modelled as a perfect cardioid [2].

## 2.5 Room acoustics

The path from an acoustic source to a receiver is described by a room impulse response. This encapsulates all reflections and other room acoustic effects. There is much research on the modelling of room impulse responses (RIRs), and Kuttruff [1979] is a comprehensive text on the subject.

There are three main methods to model room acoustic effects; analytical, geometrical and

[2]The most commonly used microphone is the Shure SM-58 which has a cardioid polar response

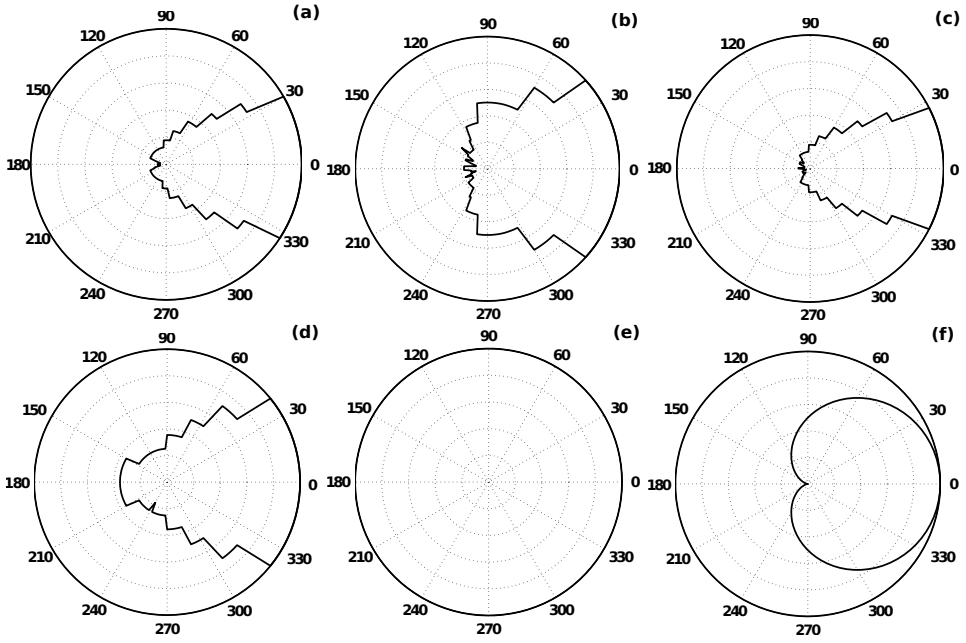


Figure 2.3: Source dispersion and receiver response patterns. a) FOH loudspeaker, based on Meyer UPA-1P @ 1kHz, b) monitor loudspeaker, based on Meyer UM-1P @ 1kHz, c) guitar amplifier, based on Meyer UPA-1P @ 2kHz, d) bass amplifier, based on Meyer USW-1P @ 250Hz, e) omnidirectional source/receiver used to model the dispersion of the vocals, all components of the drum kit and the listener response, f) cardiod pattern used to model the microphone response.

finite difference time domain (FDTD). Analytical methods solve the wave equation directly, and are only applicable to very simple geometries. Geometrical methods simplify the description of a sound wave to that of a ray by dividing a spherical sound wave emitted from a simple omnidirectional source into discrete quanta of sound energy. FDTD methods discretise the room into a large number of simple elements and numerically solve the wave equation for each element.

The geometrical method called the *Image Source Method* is used in this thesis, early work on which was presented by Allen and Berkley [1979] and Gibbs and Jones [1972]. The image source method identifies the static, virtual positions of reflected sources by plotting images of the room and source about the room boundaries. This is illustrated in Figure 2.4 in which three image rooms have been plotted. The path of a sound ray from each image source to the receiver can be traced. Each point where this path crosses a wall is a reflection point. The order of an image room is the number of times the sound ray has been reflected before reaching the receiver. The attenuation due to distance $g_d$ and the time delay $\delta t$ are evaluated from the length of the path $d$, and the wall absorption is modeled using a broadband gain $g_\alpha$.

The reduction in sound intensity as a function of $d$ follows the inverse square law. The sound pressure is proportional to the square root of the intensity, and so is inversely proportional to $d$.

This translates as a halving of the level each time the distance doubles. If $d_{ref}$ is the reference distance at which the level of an acoustic source is stated, then the change in level at distance $d$ is given by,

$$g_d = \frac{1}{2}^{\frac{log(d_{ref}/d)}{log(2)}}. \tag{2.3}$$

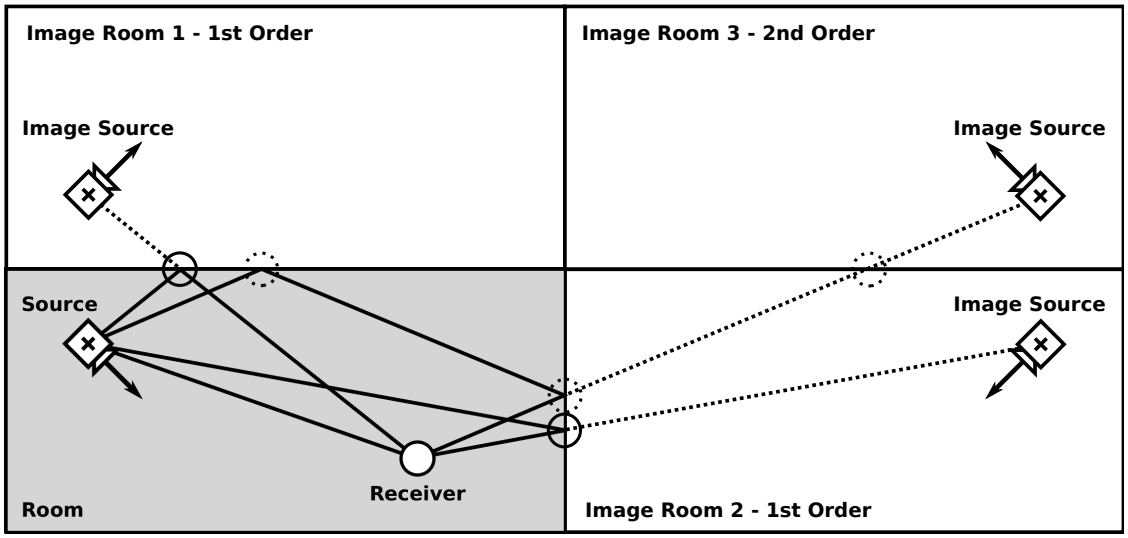


Figure 2.4: Diagram illustrating the image source method. The dashed lines emanating from the sources and image sources are the sound ray paths, and the solid lines show how these paths travel around the room. The solid and dashed circles highlight the wall crossings i.e. reflection points, for first and second-order image rooms respectively. For simplicity, only a few reflections and image sources are shown. (There is actually one first-order reflection per room wall, one second-order reflection per first-order image room wall, and so on). The room impulse is evaluated by combining the sound ray paths from all image sources.

## 2.6 The mix equation

A sound ray leaving a second order image source will reach the receiver location after a time $\delta t$ and will have a gain of $RA_R g_d g_{\alpha_1} g_{\alpha_2}$, relative to the on-axis acoustic source signal at the reference distance of 1 m, where $g_{\alpha_1}$ and $g_{\alpha_2}$ represent the absorption for two different reflection points (the subscripts $I$ and $L$ which discriminate between instruments and loudspeakers have been removed for clarity). The complete RIR is denoted by $h$, and is evaluated in the time domain by combining the sound rays from all image sources. In the frequency domain its transfer function is given by $H$, and using this, the acoustic signal from instrument $i$ at the receiver location $r$ is given by,

$$X_{R_{ir}} = X_{I0_i} \left( H_{I_{ir}} + \sum_{l=1}^{N_L} P_{il} H_{L_{lr}} \right), \tag{2.4}$$

where $i$, $l$ and $r$ are instrument, loudspeaker and receiver indices, $I$, $L$ and $R$ distinguish between instruments loudspeakers and receivers, $P$ is the processing applied in the mixing desk, and $N_L$ is the number of loudspeakers. It is a combination of the direct signal path from instrument to receiver, and the reinforced paths that have passed through the mixing desk and been output by the loudspeakers. Equation 2.4 enables the mix experienced at each receiver location to be evaluated based on the instrument on-axis acoustic signals. When applied to the mixing of recorded music, the transfer function of the direct path, $H_I$ is removed, and an additional calibration term is required that converts the audio signals into acoustic signals.

## 2.7 Model approximations

The development of Equation 2.4 required three assumptions to be made:

1. Radiation patterns of the sources and the responses of the receivers are modelled using polar broadband gain functions. This will introduce errors in the predicted mixes and estimation of acoustic feedback gain.

2. RIRs are estimated using the image source method. This does not include wavelength and phase information and means that diffraction, refraction and interference effects will be absent from the modelled RIRs.

3. Reflections are modelled as being specular, which means that all of the reflected energy is contained within a sound ray and that the angle of reflection of the sound ray is equal to the angle of incidence. Real reflections are to some extent diffuse, which means that some of the energy is spread out rather than being contained in a ray. Specular reflections are perfectly correlated with the original acoustic signal and will increase the affect of coloration and comb filtering.

All of these approximations relate to the estimation of the RIR. Assuming that the automatic mixing concept that is being proposed can be proven, it is a relatively simple task to introduce more advanced room acoustic models, or indeed to measure the RIRs from venues where the system is to be implemented. Moreover, in the case of recorded music, a single impulse response measurement from each loudspeaker to the mixing engineer will adequately describe the signal paths.

## 2.8 Mix features

The model will be used to do automatic mixing for live music, but before proceeding a number of decisions must be made with respect to the mix features. The reference features can either be generic, as with the work by Perez-Gonzales and Reiss [2007, 2008, 2009a,b, 2010], or they can be specific to a certain song, as with the work by Barchiesi and Reiss [2009, 2010], Kolasinski [2008], Terrell and Reiss [2009]. The former case represents fully-automatic mixing, whereas the latter may be considered semi-automatic, because the features used in the mixing algorithm have been extracted from a manually produced mix. The fully automated case seems to be a favourable long term option, but as yet the heuristic models used to produce a generic reference mix have not been fully validated. In addition, the objective of the live mixing work is to automate the practical issues, but to still permit some manual control over the mix. For these reasons, the semi-automatic approach is taken.

Two broad groups of automatic mixing tools were outlined in the previous chapter, which were real-time and offline. Implementing specific, real-time target mix features would necessitate a dynamic description of the reference mix and a knowledge of the temporal position within the song. Although an interesting idea, a proof of concept using an off-line approach is a sensible first step and is adopted here. The off-line approach provides a static set of control parameters, and these are derived from a specific section of audio. In this respect the approach taken is analogous to the sound check, during which the engineer will set up good, general mixes to which relatively minor adjustments are made during the performance. Application of the model will focus on providing good, baseline mixes equivalent to those produced during a sound check.

Sound features were discussed in the previous chapter. It is the intention to develop new loudness-based perceptual features for automatic mixing applications, but before this can be done, the practical issues associated with live music performance must be tackled. So long as the features used are evaluated from acoustic signals, the extension to perceptual features, once validated, is a relatively simple step. For this reason, for the next two chapters, the mix is described using objective sound features, namely the RMS level of acoustic signals that form the mix, in dBSPL.

Mix features are extracted for a section of a song (for example, a chorus) in which all instruments are active. A typical chorus in a rock band is 16 bars which at 120 beats per minute (bpm) is around 30 seconds long. Convolving signals of this length with transfer functions at

each iteration of an optimisation algorithm would result in a substantial solution time. In order to reduce the length of the sample, the incoherent average of Fast Fourier Transform (FFT) windows is taken, using a sliding normalised Hanning window, and a 50% overlap. The length of the window is set equal to the length of the room impulse response, which is 1 second at 44.1 kHz. The acoustic signal at the receiver location can be evaluated in the frequency domain using Equation 2.4. The RMS level, given by $p_{ir}$, can also be evaluated in the frequency domain using,

$$p_{ir} = \frac{1}{N}||X_{R_{ir}}|| \tag{2.5}$$

where $N$ is the number of frequency points.

The absolute RMS level of each instrument in the mix is evaluated using Equations 2.4 and 2.5, and are stored in the vecotr $m_{L_a}$, where the subscript $L$ identifies that it is a listener mix (as opposed to a microphone), and the subscript $a$ shows that it is the absolute level in dBSPL. The relative RMS level, denoted by $m_{L_r}$, is used to describe the mix, and is expressed as the differences in absolute levels, relative to an arbitrarily chosen reference instrument. For example, a mix containing instruments with absolute RMS levels of,

$$m_{L_a} = \begin{bmatrix} 80 & 78 & 83 & 76 \end{bmatrix}, \tag{2.6}$$

is described, using instrument 1 as the reference by

$$m_{L_r} = \begin{bmatrix} 0 & -2 & 3 & -4 \end{bmatrix}. \tag{2.7}$$

The levels of the second, third and fourth instruments are respectively 2 dB lower, 3 dB higher and 4 dB lower than the first instrument.

## 2.9 The objective function

In this section the formulation of the objective function is described. This includes quantification of the mix errors, definition of the control parameters, and the introduction of a constraint to prevent the onset of acoustic feedback.

### 2.9.1 Mix errors

If the reference mix, $m_{R_r}$, is described by,

$$m_{R_r} = \begin{bmatrix} 0 & -3 & 1 & -4 \end{bmatrix}, \tag{2.8}$$

then the error $e_l$ in the mix at listener location $l$ is given by,

$$e_l = ||m_{R_r} - m_{L_r}||. \tag{2.9}$$

The total error in all mixes, $\varepsilon_T$ is found by summing the mix errors at each location subject to a diagonal weighting matrix $W$, which allows priority to be assigned to the requirements of some listeners[3],

$$\varepsilon_T = eWe^T. \tag{2.10}$$

The objective of the optimization algorithm is to minimize $\varepsilon_T$. This is done by adjusting the control parameters of the audio effects within the mixing console. Each audio effect applies a linear transformation, and it is assumed that individual transformations can be applied to each instrument and loudspeaker pair. The transformations are identified by $P_{M_{il}}$ where $i$ is the instrument index and $l$ is the loudspeaker index.

### 2.9.2 Control parameters

The audio effects are restricted to broadband gains, which on a mixing console of often referred to as faders. This equates to sending linearly scaled version of each instrument's acoustic signal through each loudspeaker. Mixing consoles also have equalisation controls, and many have additional audio effects such as reverberation and compression, but these are not included here. By removing frequency dependent audio effects, Equation 2.4 can be rewritten as,

$$X_{R_{ir}} = X_{I0_i}\left(H_{I_{ir}} + \sum_{l=1}^{N_L} g_{C_{il}} H_{L_{rl}}\right), \tag{2.11}$$

where $g_{C_{il}}$ is the gain applied to the audio signal from instrument $i$ before output via loud-

[3]The weighting matrix $W$ is introduced to give some flexibility in how the mixes are set. Emphasis may be placed on the requirements of the audience over the performers, and for cases where the vocal performance is important, the requirements of the vocalist would most probably supersede those of the other performers.

speaker $l$. It is worth noting that Equation 2.4 can be use to implement any linear audio transformation. This includes equalisation and reverberation, but dynamic audio effects such as compressors that are non-linear, would require adaptation of the model if they were to be included.

### 2.9.3 Acoustic feedback constraint

A review of acoustic feedback control by van Waterschoot and Moonen [2011] contains a detailed explanation of acoustic feedback. Acoustic feedback results from the coupling between loudspeakers and microphones which can be described by a feedback loop transfer function, $F$, which is given by,

$$F = \sum_{m=1}^{N_M} \sum_{l=1}^{N_L} g_{C_{ml}} H_{L_{ml}}, \tag{2.12}$$

where $m$ is the microphone index. Acoustic feedback will occur if the feedback loop transfer function is unstable. The Nyquist stability criterion [Nyquist, 1932] states that a closed-loop system is unstable if the loop gain is greater than or equal to unity, or if the loop phase is a multiple of a complete cycle, at any frequency. We constrain the mixing controls to fulfill these criteria for the vocal microphone only (experience shows that vocal microphones are the most likely cause of acoustic feedback). During a performance it is possible that the feedback transfer function can change, for example if the vocalist changes position on stage with the microphone in hand. In order to account for this a buffer of 3 dB is used, so the feedback constraint is given by $f < -3$ dB, where $f$ is the maximum magnitude of $F$. The review paper [van Waterschoot and Moonen, 2011] also describes state of the art feedback control techniques. Many of these are online methods which use additional signal processing of the microphone signals to disrupt coupling between the loudspeakers and microphones to satisfy the Nyquist stability criteria. Such techniques are not considered here.

## 2.10 Summary

In this chapter a general model of mixing has been presented that applies to both live and recorded music. The model allows the mix to be evaluated as a set of acoustic signals from each of its component sources. A focus has been placed on live mixing, where the model is more complex, due to: the presence of live acoustic sources, multiple receiver locations, and the potential for acoustic feedback. A second model, that describes the role of the engineer as a constrained, multi-

objective optimisation problem has also been developed. The errors in the optimisation problem correspond to the differences between the target and reference mixes, and the parameters are the controls on the mixing desk.

A number of decisions have been made on the nature of the mix features used, however, they do not restrict application of the mix model and automatic mixing framework. These include: the use of specific, as opposed to generic, reference mix features; the application to offline automatic mixing; the use of objective, as opposed to perceptual, mix features (namely the relative level of the acoustic signals in the mix); and the simplification of the mixing console to use only broadband gain controls. In the next chapter a live performance case study is presented that demonstrates the automatic mixing approach, and the computational issues associated with solving the optimisation problem are addressed.

# Chapter 3

# Live Automatic Mixing Case Study

---

In this chapter, the models described in the previous chapter are used to do automatic mixing in a virtual live performance. The venue is chosen to be typical of a medium performance, and the layout of the sources and loudspeakers are set accordingly. The acoustic signals generated by the instruments are simulated using pre-recorded audio signals from an unsigned band. Reference mixes are produced subjectively for each performer, and for the audience, from which the relative RMS sound levels are extracted. The optimisation algorithm is used to automatically set the gain controls on the mixing console to produce these mixes. Two optimisation strategies are trialled. The first attempts to solve the optimisation problem in one go, whilst the second splits it into two stages, and is shown to give a more robust solution in a shorter time. Following this, the amount of coupling between FOH and monitor mixes is quantified, and its effect on the ability to meet the mix requirements is discussed.

## 3.1 Virtual live performance

The case study is of a virtual performance with the setup shown in Figure 3.1. The performance is by an ensemble consisting of four performers with instruments: vocals, electric guitar, electric bass, and drum kit, which is composed of a kick drum, a snare drum, hi-hats and a cymbal. The performance takes place in a rectangular venue with dimensions 30m by 20m. The FOH mixes are evaluated at six audience locations, and the monitor mixes at each of the performer locations. The sound reinforcement system consists of two FOH loudspeakers and four monitor loudspeakers. The wall behind the performers was modelled as having an absorption coefficient

of 0.7 (heavy curtain at 2 kHz) and the other walls with an absorption coefficient of 0.4 (course concrete block at 2 kHz). Absorption coefficients were taken from [Tontechnik-Rechner, 2012]. This represents additional acoustic damping such as heavy curtains that are commonly used at the back of the stage.

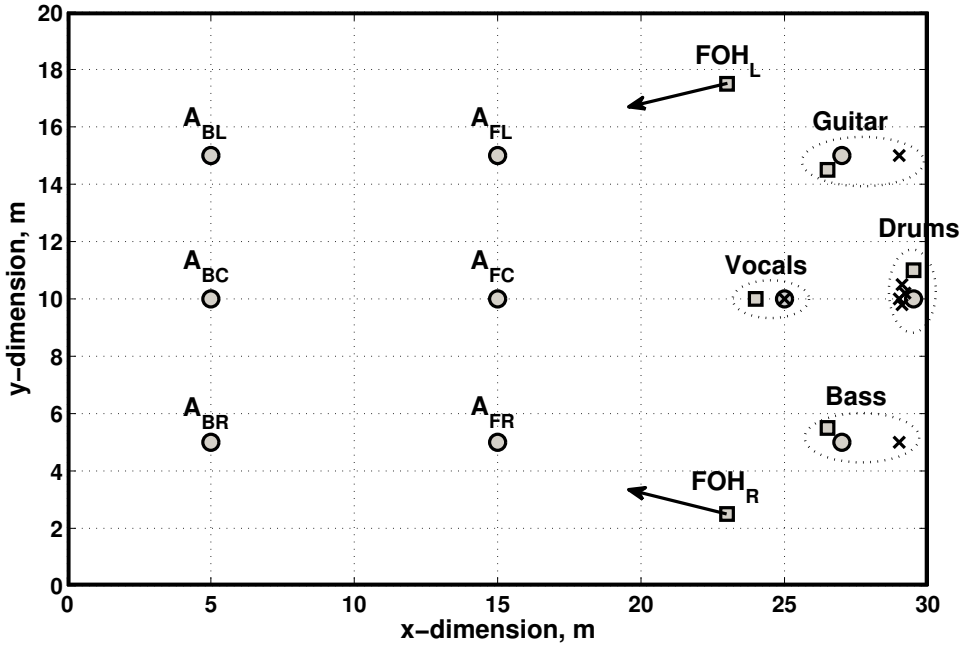


Figure 3.1: Diagram showing the layout of the venue. Listeners are identified by ○, instruments by ×, and loudspeakers by □. Audience locations face the performers and are labeled using the convention $A_{XY}$, where $X$ is: $B$ for back or $F$ for front, and $Y$ is: $L$, $C$ or $R$ for left, centre or right respectively. The performers, their instruments, and monitor loudspeakers are grouped and labelled. For guitar and bass the instrument location is the amplifier; for the vocals, the instrument and performer locations coincide. Performers and instruments face the audience and each monitor loudspeaker faces the performer to whom it is assigned. The orientation of the FOH loudspeakers are identified by arrows.

### 3.1.1 Acoustic signals and reference mixes

The chorus of a multi-track recording of an unsigned band is used to generate the acoustic signals. For each instrument, the corresponding track from the audio recording is scaled to give realistic peak and RMS levels for on axis reference distances of 1m. These are shown in Table 3.1. The waveforms of the acoustic signals, plotted in terms of absolute pressure (Pascals), are shown in Figure 3.2.

Reference mixes were produced subjectively using a digital audio workstation (DAW) for the audience and for each performer. The intention is for the live automatic mixing system to be

| | Vocals | Guitar | Bass | Kick | Snare | Hi-Hat | Cymbal | Mix |
|---|---|---|---|---|---|---|---|---|
| $p_{RMS}$ (dBSPL) | 61.0 | 77.0 | 85.0 | 81.0 | 72.0 | 45.0 | 61.0 | 87.1 |
| $p_{peak}$ (dBSPL) | 72.7 | 90.0 | 91.6 | 95.7 | 83.3 | 65.5 | 79.7 | 100.9 |

Table 3.1: The peak and RMS SPLs in dB SPL of the acoustic signals, on-axis, at a reference distance of 1m.

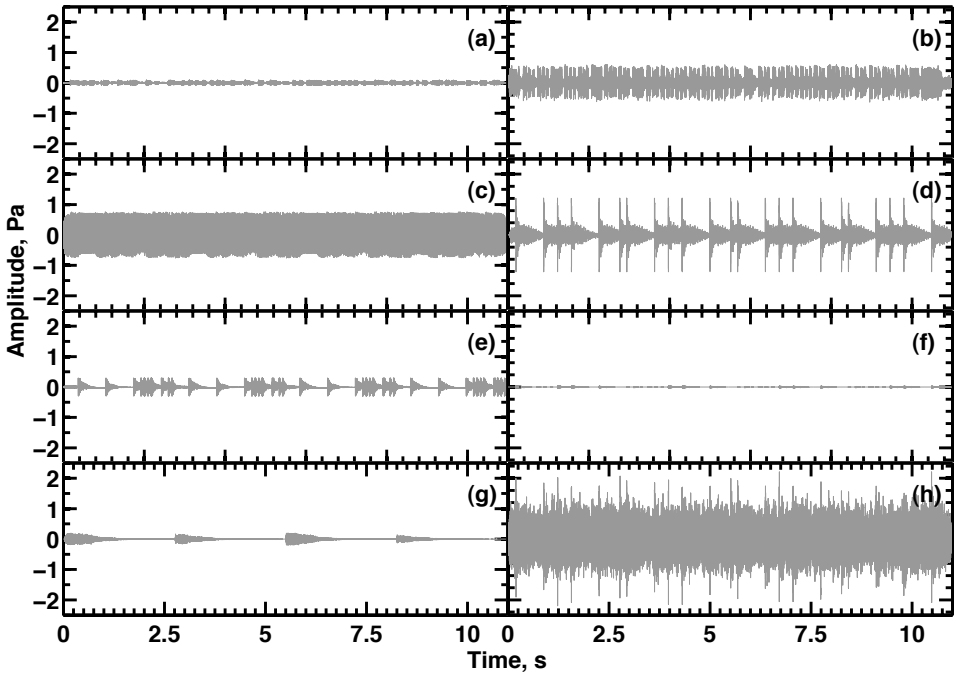


Figure 3.2: The on-axis acoustic instrument signals, plotted in terms of absolute pressure (Pascals), where (a)-(h) corresponds to vocals, guitar, bass, kick drum, snare drum, hi-hat and cymbal. RMS and peak levels in dBSPL are given in Table 3.1

.

used in this way by non-expert engineers, i.e., they will be able to produce their mixes using a recorded version of the song, from which the reference mix features are extracted. The automatic system then takes care of the practical issues associated with live performance.

The reference mix features, $m_{R_r}$, are shown in Equation 3.1, where rows 1 to 5 correspond to: the vocalist, the guitarist, the bassist, the drummer and the audience respectively; and columns 1 to 7 correspond to: vocals, guitar, bass, kick drum, snare drum, hi-hats and cymbal respectively. This ordering convention is maintained throughout this case study.

$$m_{R_r} = \begin{bmatrix} 0 & -4.2 & 2.6 & 0.4 & 0.5 & -18.2 & -14 \\ 0 & 1.7 & 5.2 & 2.7 & 2.6 & -16.2 & -11.9 \\ 0 & -3.7 & 8.4 & 8.5 & 6.8 & -16.2 & -11.9 \\ 0 & -2.3 & 7.7 & 6.5 & 6.1 & -14.6 & -12.6 \\ 0 & -0.4 & 6.7 & 2.7 & 2.8 & -18.6 & -11.8 \end{bmatrix} \tag{3.1}$$

### 3.1.2 Additional constraints and considerations

There are 7 instruments and 6 loudspeakers in the virtual performance, which gives 42 control parameters in total. An additional constraint was introduced to make all instruments centrally panned in the FOH mix, i.e. the same signal is sent to the left and right FOH loudspeakers. The model of the listeners does include panning perception, and would not be able to differentiate between a hard pan to left or right, or centre panning. This constraint prevents mixes with extreme or undesirable panning settings, although future work should incorporate panning effects. This leaves 35 independent control parameters.

The control parameters are optimised using a combination of a gradient descent search method and a genetic algorithm, implemented using the Matlab *fmincon* and *ga* functions respectively. Gradient descent search methods calculate local gradients in the error function with respect to the control parameters, and use these to determine the search direction for the next iteration. These methods are deterministic, and are unable to escape local minima. Genetic algorithms begin with a random population of solutions, and through an evolutionary process, including random mutation, the solution converges to the minimum. These methods are quasi-random, and are able to navigate a solution space with many local minima.

In a live performance the most important mix is that heard by the audience. However, in order for the performers to perform to their best ability they must be able to at least hear what they are playing. This is especially true for vocalists, some of whom may be are unable to sing in tune without sufficient feedback from the monitoring system. For these reasons, the listener's requirements are weighted such that those of the vocalist and the audience are twice as important as those of the other performers [1], i.e.,

$$W = [diag(2,1,1,1,2,2,2,2,2,2)]. \tag{3.2}$$

## 3.2 Optimisation strategies

At this point, the role of the mixing engineer has been defined as an optimisation problem based on the following: a model to evaluate the mix at each location (Eqn 2.4), a set of features that describe the mix (Eqn. 2.7) , a means to quantify the error in the mixes relative to the reference (Eqn. 2.10), a set of control parameters that can be used to modify the mixes ($g_C$ in Eqn. 2.11),

[1] The values were chosen based on the authors opinions of relative importance in mix requirements.

and a constraint that prevents control parameter settings that cause acoustic feedback (Eqn. 2.12). The next stage is to solve this optimisation problem using the algorithms discussed in Section 3.1.2.

### 3.2.1 Brute force optimisation

The first attempt at solving the optimisation problem used a brute force approach, i.e. the error and constraint functions were defined, the initial control parameters were set, and were collectively passed to an optimisation function (i.e *fmincon* or *ga*), which attempted to minimise the error. This approach was trialled with different sets of optimisation parameters. For the gradient descent method the number of iterations was varied from 5-50, and for the genetic algorithm the population size was also varied from 5-50. The residual errors are plotted against optimisation time in Figure 3.3(a). For the gradient descent search method, identified by $\times$, the starting solution set all gain controls to $-\infty$ dB, and very little improvement upon this is found even for larger numbers of iterations, i.e. the solution gets stuck in a local minima very quickly. The genetic algorithm, identified by $+$, faired better, and was improved further by using its optimal solution as the starting point for a subsequent gradient descent search stage, identified by $\bigcirc$, however, the performance of this combined algorithm is difficult to judge because it is not known whether it represents the global minimum.

### 3.2.2 Algorithm limitations

The algorithms can be improved through consideration of the optimisation problem. There are two key issues: linear dependency between control parameters, and the feedback constraint. There is strong linear dependency between the control parameters. This can be understood by considering a situation where the SPL of the guitar must be decreased relative to the bass. The engineer can either turn the guitar down or turn the bass up, and he can do this on any loudspeaker or combination of loudspeakers. This results in a solution space with many local minima, and as seen in Section 3.2.1, the gradient descent method gets stuck. The feedback constraint restricts the feasible solution space. Gradient descent methods can become stuck on a constraint boundary, and genetic algorithms are slowed significantly, because many more function evaluations are required at each generation to ensure that all solutions in the subsequent population are feasible.

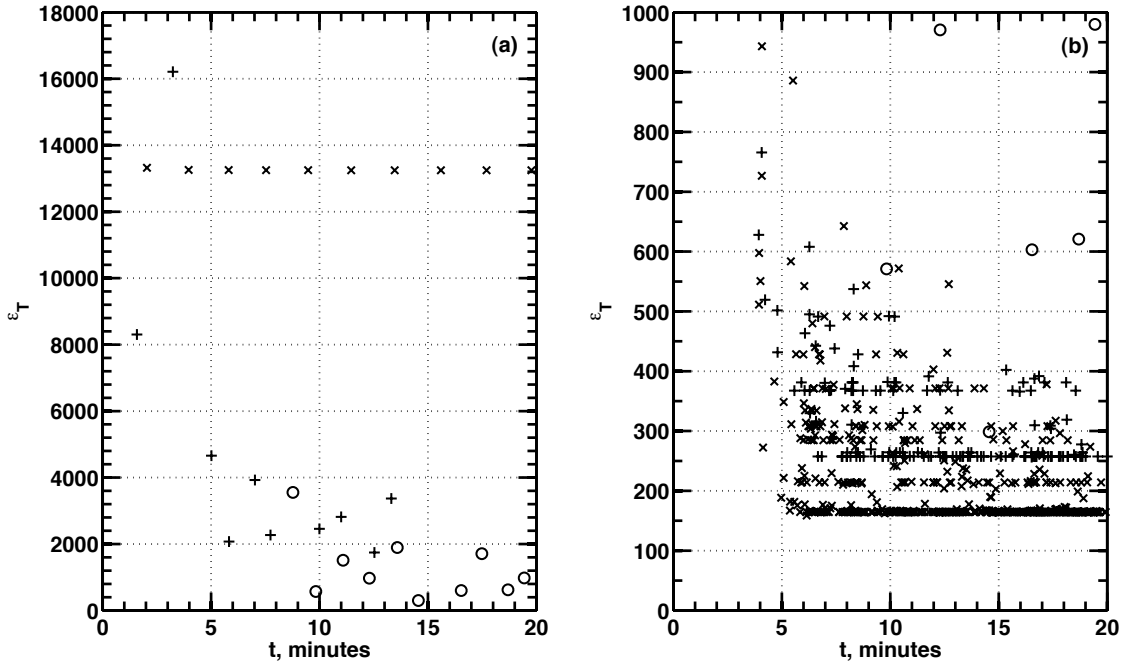


Figure 3.3: The residual of the error function plotted against solution time. Part a) shows results for the direct approach in which: $\times$ uses the gradient descent search, $+$ uses the genetic algorithm and $\circ$ uses the genetic algorithm followed by the gradient descent search. Part b) shows results for the targeted approach in which $+$ are members of cluster 1 and $\times$ are members of cluster 2. The direct approach using the combined optimisation algorithms are also plotted and are identified by $\bigcirc$.

### 3.2.3 Targeted optimisation

The genetic algorithm performs better because it can deal with local minima; however, solving the entire optimisation problem in one go with a genetic algorithm is inefficient, because of the increased time penalty when enforcing the feedback constraint. The feedback constraint is a function of the vocal control parameters only (it has been assumed that the vocal microphone is the most likely cause of acoustic feedback), and by separating the optimisation into two parts, the first of which sets the vocal control parameters only, the dimensionality of the constrained optimisation problem can be reduced, enabling a more thorough search of the solution space in a shorter time. Once the vocal control parameters have been set, those of the other instruments can be found quickly as they are no longer constrained. Separating the optimisation in this way also reduces the effect of linear dependency because the control parameters of the remaining instruments are set relative to a fixed vocal level.

An objective function is required to set the vocal control parameters. It is not possible to use the existing function because it delivers features of the reference mixes as opposed to maximising

the vocal level. There are many ways to maximise the vocal level, but a comparison is not made here. The chosen objective is to make the vocal RMS SPL at each listener location higher than required with respect to all other instruments in the mix, when considering the direct sound only. The vocal error for listener $l$ is defined by,

$$e_{v_l} = ||MAX\,(0, m_{L_{il}} - m_{R_{il}})\,||, \tag{3.3}$$

where $i$ is the instrument index, $l$ is the listener location, and where the vocal is chosen as the reference instrument when calculating the relative levels (see Eqn. 2.7). The total vocal error, $\varepsilon_{v_T}$ is evaluated using the weighting matrix as for $\varepsilon_T$ in Equation 2.10,

$$\varepsilon_{v_T} = e_v W e_v^T. \tag{3.4}$$

To summarise, the vocal gain controls are set by minimising $\varepsilon_{v_T}$ using Equation 3.4, then the gain controls for the other instruments are set by minimising $\varepsilon_T$ using Equation 2.10. In both stages of the optimisation process, the genetic algorithm is used, followed by the gradient descent method, which uses the output of the genetic algorithm as its starting solution. To determine suitable optimisation parameters that give a reliable solution, the optimisation is run with: genetic algorithm population sizes from 5 to 50, followed by gradient descent search method iterations of 5 to 50; and the residual errors and optimisation times are compared. For each combination of algorithm parameters, 10 solutions are obtained to account for the quasi-random nature of the genetic algorithm, and the residuals are plotted against optimisation time in Figures 3.3 (b).

### 3.2.4 Multiple solutions and clustering

It is clear from Figure 3.3 that the targeted optimisation strategy far out-performs the brute force (note the different scales on the y-axis). The minimum solution is $\varepsilon_T = 158.8$, and although it cannot be proved definitively, it appears to be the global minima. The consistency in the solutions is investigated using a k-means clustering method. This is an established machine learning technique further details of which can be found in Witten and Frank [2005]. The number of clusters was estimated using Ward's method [Ward, 1963], which clearly indicated two clusters. In Figure 3.3 (b) solutions belonging to clusters 1 and 2 are indicated by $+$ and $\times$ respectively. Also plotted using $\bigcirc$ are the solutions from the brute force strategy. The correlation of each solution with the centroid of its respective cluster is over 95% for all solutions for which the genetic

algorithm population size is 10 or more.

The control parameters and residual errors corresponding to the best solution in each cluster are shown in Table 1. The gain controls generally show significant vocal amplification, and there is minimal amplification (via microphones and loudspeakers) of guitar, bass and kick drum, which demonstrates the contribution of the direct signal path to the mixes. Both clusters have small residuals at the audience and vocalist locations and larger residuals at the other performer locations, in particular the drummer location. This is attributed to the weightings used and the increased contribution of the direct sound of these performers own instruments at their respective locations. This can be seen by inspecting the individual instrument errors. The largest errors for the guitarist, bassist and drummer correspond to their own instrument, the levels of which are substantially higher than required (a positive error means the instrument level is too high compared to the vocals).

The main differences between the two clusters is the vocal gain sent to the monitor loudspeakers. Cluster 2 shows a relatively even spread of vocal gain across the monitor loudspeakers, whereas cluster 1 shows a high concentration of vocal gain output via monitor loudspeaker 1 (the vocalist's monitor), and minimal vocal gain output via monitor loudspeaker 4 (the drummer's monitor). Figures 3.4(a) and (b) are plots of the absolute RMS level of the vocal signals in dB-SPL, throughout the room for clusters 1 and 2 respectively. To reduce the computational time the vocal level was evaluated without including room reflections. The RMS level at the vocalist location is around 8 dB higher in cluster 1 compared to cluster 2, and because the mix errors are comparable, this means that the levels of all other instruments must also be 8 dB higher. Figure 3.4(a) shows that the acoustic signal from the vocal monitor loudspeaker is providing substantial vocal SPL to the drummer location to compensate for the minimal vocal SPL from the drum monitor.

Which of these solutions is best is debatable. Cluster 2 shows a lower residual and gives a more even sound level on stage, whereas cluster 1 indicates that it may be possible to remove the drum monitor and still obtain a reasonable mix at the drummer's location. Either solution is acceptable and can be found reliably using the targeted approach. The difference between the two solutions arises from the vocal gain settings that are minimised in the first stage of the algorithm. It would be possible to add further constraints to the objective function which favour one solution, or to increase the population size or number of iterations to ensure that

| | | | **Vocals** | **Guitar** | **Bass** | **Kick** | **Snare** | **Hi-Hat** | **Cymbal** | $e_R$ |
|---|---|---|---|---|---|---|---|---|---|---|
| **Cluster 1** | Gain | FOH L | 15.5 | -2.8 | -4.0 | -8.2 | 1.9 | 12.5 | 5.4 | |
| | | FOH R | 15.5 | -2.8 | -4.0 | -8.2 | 1.9 | 12.5 | 5.4 | |
| | | Mon V | 17.6 | -2.7 | -4.9 | -4.6 | 4.3 | 15.0 | 5.1 | |
| | | Mon G | 3.9 | -24.9 | -16.5 | $-\infty$ | -8.6 | 4.1 | -6.1 | |
| | | Mon B | 4.0 | -16.3 | -13.2 | -12.0 | -1.5 | 4.3 | -5.9 | |
| | | Mon D | -72.9 | -21.6 | -15.8 | $-\infty$ | -7.7 | -2.8 | $-\infty$ | |
| | Error | Vocalist | 0 | -0.6 | 0 | 0.4 | 0.4 | 0 | 0.4 | 0.9 |
| | | Guitarist | 0 | -3.3 | 0.1 | -1.7 | 0.1 | 0 | 0 | 3.7 |
| | | Bassist | 0 | -0.3 | -3.2 | 0.2 | 0.2 | 0 | 0 | 3.3 |
| | | Drummer | 0 | -0.3 | 0.2 | -11.7 | -2.2 | 0 | -7.7 | 14.2 |
| | | Aud FL | 0 | -0.3 | -0.2 | -0.1 | 0.6 | 0 | 0 | 0.7 |
| | | Aud FC | 0 | 0.3 | 1.4 | 1.1 | 0.3 | 0 | 0.1 | 1.8 |
| | | Aud FR | 0 | 0.9 | -0.8 | -0.4 | 0.3 | -0.1 | 0.1 | 1.3 |
| | | Aud BL | 0 | -0.4 | -1.3 | -1.4 | -1.1 | 0.1 | -0.2 | 2.3 |
| | | Aud BC | 0 | -0.1 | -0.3 | 0 | 0.5 | 0 | 0.2 | 0.6 |
| | | Aud BR | 0 | -0.2 | 0.1 | -1.3 | -1.1 | 0.1 | -0.1 | 1.7 |
| | | $\varepsilon_T$ | | | | | | | | **256.3** |
| **Cluster 2** | Gain | FOH L | 15.8 | -2.4 | -4.0 | -7.8 | 3.4 | 12.9 | 5.6 | |
| | | FOH R | 15.8 | -2.4 | -4.0 | -7.8 | 3.4 | 12.9 | 5.6 | |
| | | Mon V | 8.9 | -12.2 | -14.6 | -28.8 | -6.2 | 6.4 | -3.5 | |
| | | Mon G | 4.2 | -24.1 | -15.9 | -48.5 | -7.5 | 4.0 | -6.1 | |
| | | Mon B | 5.5 | -15.0 | -11.7 | -9.9 | -0.5 | 5.2 | -4.7 | |
| | | Mon D | 5.5 | -13.6 | -11.2 | $-\infty$ | $-\infty$ | 5.4 | $-\infty$ | |
| | Error | Vocalist | 0 | 0 | 0 | -0.2 | 0 | 0 | 0.1 | 0.2 |
| | | Guitarist | 0 | -3.9 | 0.1 | 0 | 0 | 0 | 0 | 3.9 |
| | | Bassist | 0 | 0 | -2.8 | 0.2 | 0.1 | 0 | 0 | 2.9 |
| | | Drummer | 0 | 0.1 | -0.1 | -9.2 | -0.2 | 0 | -4.9 | 10.4 |
| | | Aud FL | 0 | -0.4 | -0.2 | -0.3 | 0.4 | -0.1 | 0 | 0.6 |
| | | Aud FC | 0 | 0.2 | 1.5 | 0.9 | 0 | 0 | 0.1 | 1.8 |
| | | Aud FR | 0 | 0.8 | -0.7 | -0.6 | 0.4 | -0.1 | 0.2 | 1.3 |
| | | Aude BL | 0 | -0.4 | -1.2 | -1.5 | -0.8 | 0.1 | -0.1 | 2.1 |
| | | Aud BC | 0 | -0.3 | -0.4 | 0.5 | 0.5 | -0.1 | 0.1 | 0.8 |
| | | Aud BR | 0 | -0.1 | 0 | -1.5 | -0.7 | 0.1 | -0.2 | 1.7 |
| | | $\varepsilon_T$ | | | | | | | | **158.8** |

Table 3.2: Control parameters (gains) and residual errors for the optimial solution in each identified cluster. The gain section shows the gain applied to each instrument in the indirect path via each loudspeakers in dB. The error section shows the relative error of each instrument, the combined error at each listener location, $e_R$ and the total error for all listener locations combined, $\varepsilon_T$.

the lower residual corresponding to cluster 2 was found. However, the objective was to find a good approximation to the global minima in an acceptable solution time, and this has been accomplished, so these advancements are not considered at this stage.

## 3.3 Mix coupling

The effect of mix coupling is examined by producing independent FOH mixes using only the FOH loudspeakers, and independent monitor mixes using only the monitor loudspeakers. They are referred to as independent, because they have been set without considering how they might interact. The gain controls from the independent mixes are then combined, to produce a set of mixes that have essentially ignored coupling between FOH and monitor mixes. These are referred to as the uncoupled mixes. The independent and uncoupled mixes are compared with the cluster

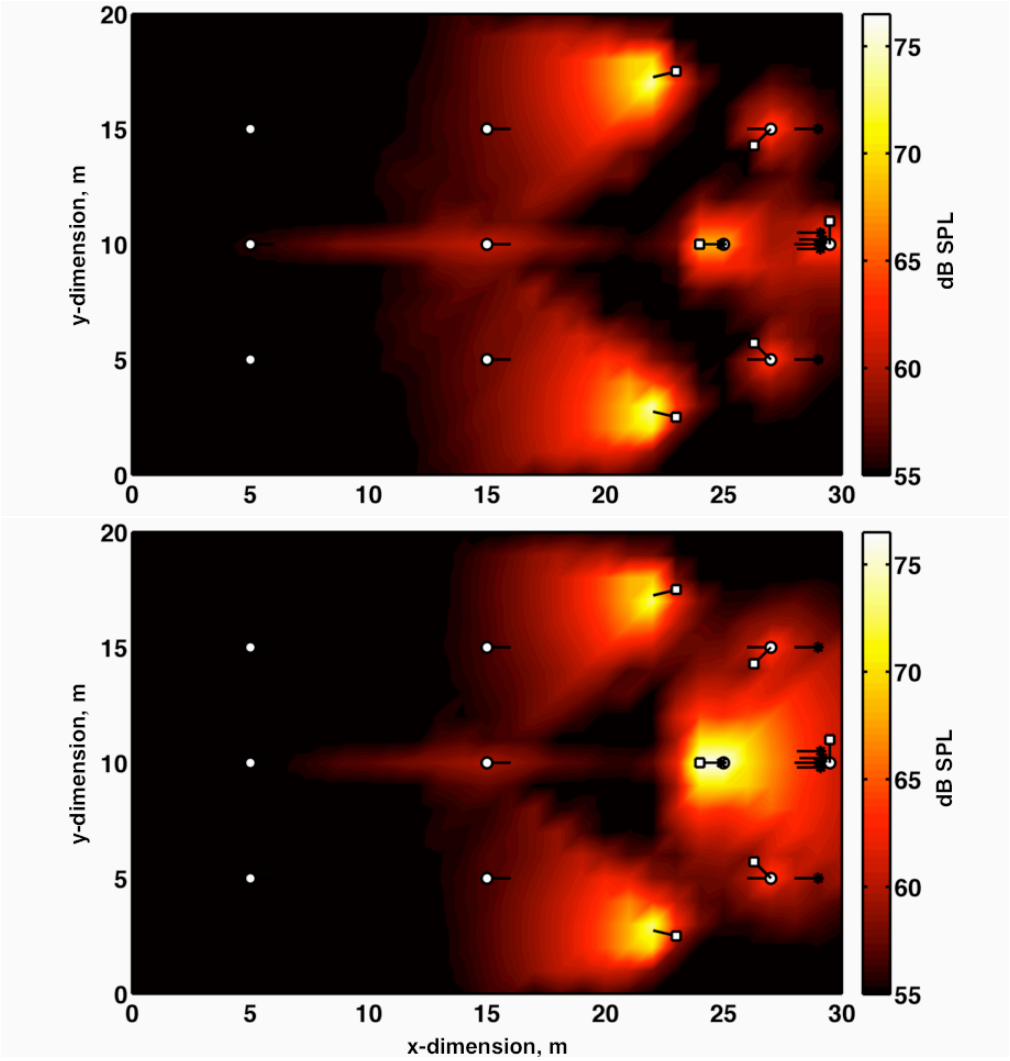


Figure 3.4: The absolute vocal sound pressure level for the two solution clusters identified in dBSPL. The top figure is cluster 1 and the bottom is cluster 2. In cluster 1, the vocalist's monitor loudspeaker produces more vocal sound energy, and its radiation pattern encompasses the drummer's location.

2 mix from the previous section. The two stage optimisation strategy is used to produce the independent mixes, with a genetic algorithm population size of 10 and a gradient descent method with 10 iterations. Table 1 shows the mix error $e_R$ at each listener location, the total error $\varepsilon_T$ and the peak feedback transfer function $f$ for the independent FOH and monitor mixes, the uncoupled mix and the coupled mix. Differences between the mix errors of the independent and uncoupled mixes indicates the level of coupling (if there was no coupling in the model then the mixes in these two cases would be indentical).

The differences for the monitor mixes are minimal except for the vocalist. The differences for the FOH mixes are slightly larger, with a maximum change of 1.5 dB at $A_{BR}$. This occurs

because the audience is relatively closer to the monitor loudspeakers than the performers are to the FOH loudspeakers, when compared to the main source of sound reinforcement (FOH and monitor loudspeakers for audience and performers respectively). The SPL of an acoustic source decreases with the distance from the source so the contribution of the monitor loudspeakers to the FOH mixes is greater than the contribution of the FOH loudspeakers to the monitor mixes. This effect is more pronounced for members of the audience at the back of the venue and this is illustrated in larger differences between FOH and uncoupled mix errors as shown in Table 3.3. The vocalist is closer to the FOH loudspeakers than the other performers so the vocalist mix is affected by the FOH loudspeaker signals. The performer mixes also have a larger contribution from the direct signal path which is independent of the sound reenforcement system.

| Mix | $V$ | $G$ | $B$ | $D$ | $A_{FL}$ | $A_{FC}$ | $A_{FR}$ | $A_{BL}$ | $A_{BC}$ | $A_{BR}$ | $\varepsilon_T$ | $f$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FOH | - | - | - | - | 0.5 | 1.2 | 1.1 | 1.6 | 0.4 | 1 | 13.7 | -8.9 |
| Monitor | 0.5 | 2.7 | 2 | 7.6 | - | - | - | - | - | - | 69.1 | -3.0 |
| Uncoupled | 1.0 | 2.6 | 1.9 | 7.6 | 0.6 | 1.9 | 1.5 | 2.4 | 1.3 | 2.5 | 108.5 | -1.3 |
| Coupled | 0.2 | 3.9 | 2.9 | 10.4 | 0.6 | 1.8 | 1.3 | 2.1 | 0.8 | 1.7 | 158.8 | -3.0 |

Table 3.3: A comparison of the residual in the error function and the feedback loop gain at each listener location when the FOH and monitor mixes are treated as being coupled.

The coupled mix errors are smaller than the uncoupled mix errors for the vocalist and the audience but are larger for the other performers, in particular the drummer. The FOH mixes contain a significant contribution from the monitor loudspeaker signals, so by considering FOH and monitor loudspeakers simultaneously improved mixes are obtained. This is also true for the vocalist as discussed previously. The large differences in the guitarist, bassist and drummer errors give an overall residual which is lower in the uncoupled case. However, inspection of Table 1 shows that $f$ is $-1.3$ dB for the uncoupled case, which is above the maximum allowable value. The difference between the guitarist, bassist and drummer mixes lie mainly in the level of the vocals, which is higher in the uncoupled case but has been obtained by breaching the feedback constraint. Table 3.3 shows that for the independent FOH and monitor mixes $f$ is $-8.9$ dB and $-3.0$ dB respectively. Setting the monitor mix in isolation pushes the feedback constraint to the limit. Any further amplification of the vocals via the FOH loudspeakers breaches the feedback constraint.

The coupling between mixes can be considered in two ways. The first is simply the contribution of both sets (FOH and monitor) of loudspeakers on the mix at a given location. It has been shown that the FOH and vocalist mixes are affected by both sets and that the monitor mixes

of the other performers are affected minimally by the FOH loudspeakers. The result of this is improved mixes at FOH and vocalist locations when coupling is taken into account. The second is the effect of the feedback constraint which must be considered for the whole sound reenforcement system. The gain that can be applied to the vocal signal can be viewed as a resource that must be shared between FOH and monitor mixes. If set independently it is likely that one set of mixes will use up all of this resource leaving none for the other. This was shown to happen with the monitor mixes which are clearly very sensitive to the feedback constraint.

## 3.4 Summary

The live performance model developed in the previous chapters has been used to do automatic mixing at a virtual live performance. The system can automatically set the gain controls on a mixing console to deliver the features of multiple reference mixes, to multiple listener locations during a live performance, whilst preventing the onset of acoustic feedback. The reference mixes cannot be obtained exactly due to the coupling between mixes, and the restrictive nature of the feedback constraint, so a best fit to all requirements is found. Through consideration of the mechanics of the optimisation algorithm, a two stage optimisation process has been developed that sets the vocal gain controls before those of the other instruments, which out-performs a direct, brute force approach.

The system takes as input a reference mix, which can be produced using a standard mixing console for recorded music. Once input, the system automatically deals with the practical issues associated with live music discussed above, and delivers a best fit to the reference mix. The complexity of live mixing has therefore been reduced to the same level as recorded mixing, and goes some way to unifying the two forms of mixing.

This chapter has demonstrated live automatic mixing for a single venue size, chosen to be representative of a medium venue. It was argued in Chapter 2 that mix coupling and the effect of acoustic feedback are more prevalent in small venues. These effects have been demonstrated in this chapter, but they have not been quantified in relation to the size of the venue. The following chapter applies the optimisation process to a range of venue sizes to determine the relationship between venue size, acoustic feedback, and coupling.

# Chapter 4

# The Effect of Venue Size on Automatic Mixing

---

The case study in the previous chapter demonstrated automatic mixing on a virtual live performance, and the effect of coupling between mixes, and the contribution of direct sound were shown. The degree to which these phenomena affect the mixing process is expected to be dependent upon the size of the venue, as discussed in Chapter 1. For automatic applications in large venues they are ignored [Perez-Gonzales and Reiss, 2009a,b], and in small venues they are expected to be more significant, however, no data are available. In this chapter, the automatic mixing algorithm is applied to venues of different sizes to quantity the relationship between the venue size and these phenomena, and their effect on the mixing process.

## 4.1 Model parameters

The venue from the previous chapter (Fig. 3.1) is used here, and is referred to as the baseline venue. The baseline venue is transformed into venues of different sizes, and in each, the algorithm developed in the previous chapter is used to do automatic mixing. The dimensions of the baseline venue, and the $(x,y)$ coordinates of the listeners are scaled linearly by factors of $2^d$, where $d$ ranges from -2 to 2 in increments of 0.5. This gives nine venues ranging from one quarter (7.5 m by 5 m) to four times (120 m by 80 m) the original size. The $(x,y)$ coordinates of the loudspeakers and instruments are also scaled linearly with two exceptions. Firstly, the monitor loudspeakers are positioned 1 m from, and facing, their respective performer, and secondly, the position of each component of the drum kit is fixed, relative to the drummer location. The absorption coefficients and mix error weighting matrix (see Eqn. 3.2) used in Chapter 3 are

| | Vocals | Guitar | Bass | Kick | Snare | Hi-Hat | Cymbal | Mix |
|---|---|---|---|---|---|---|---|---|
| $SPL_{RMS}$ | 90.0 | 102.0 | 113.4 | 115.2 | 103.7 | 92.2 | 89.3 | 117.7 |
| $SPL_{peak}$ | 108.5 | 110.8 | 122.8 | 128.4 | 124.3 | 117.2 | 116.1 | 131.7 |

Table 4.1: The sound pressure levels of the acoustic signals at a reference distance of 1m, on-axis, in dB SPL.

retained.

The levels of the acoustic sources used in the previous chapter proved to be underestimates. Whilst they were still sufficient to demonstrate the algorithm, more realistic values are required, particularly to accurately account for the contribution of direct sound to the mixes. As a result, new audio recordings are used (the set of instruments was the same), and the sound levels of the seven acoustic sources, i.e. voice, electric guitar, electric bass, kick drum, snare drum, hi-hats and cymbal were measured, as follows. A 1 kHz sine wave tone was generated using a loudspeaker, and was recorded onto a digital audio workstation (DAW), using a reference microphone with a cardioid response (DPA-4011A) and a Tascam US-122L audio interface. A sound pressure level (SPL) meter was placed at the microphone location and the rms level of the tone was recorded (it was 75 dB SPL). The same microphone, with unchanged interface settings, was used to record the acoustic sources at a reference distance of 1m for voice and guitar, and 2m for bass and the drum components (the larger reference distance was needed for the bass and drum components to account for their high peak levels). Through comparison with the recorded reference tone, the peak and mean levels of the sources at 1m reference distances were determined (sources recorded at 2m were converted to a 1m reference distance by adding 6 dB), and are given in Table 4.1. The scaled acoustic signals are plotted in Figure 4.1, with units in Pascals (Pa).

## 4.2 Overall mix level

Up until this point, no mention has been made of the overall (absolute) level of the mix, and at present, there is no means to control it, because both stages of the optimisation algorithm deal with relative levels. Initial testing of the algorithm for larger venue sizes resulted in some very low mix levels. To demonstrate this, the automatic mixing algorithm is used to set mixes for the front centre audience location[1] for all venue sizes. The overall mix level is plotted in Figure 4.2

[1]This location is chosen because it is closest to the stage and so naturally has the highest overall level amongst the audience locations, and by considering no other mixes it is more likely that the reference mix will be produced.

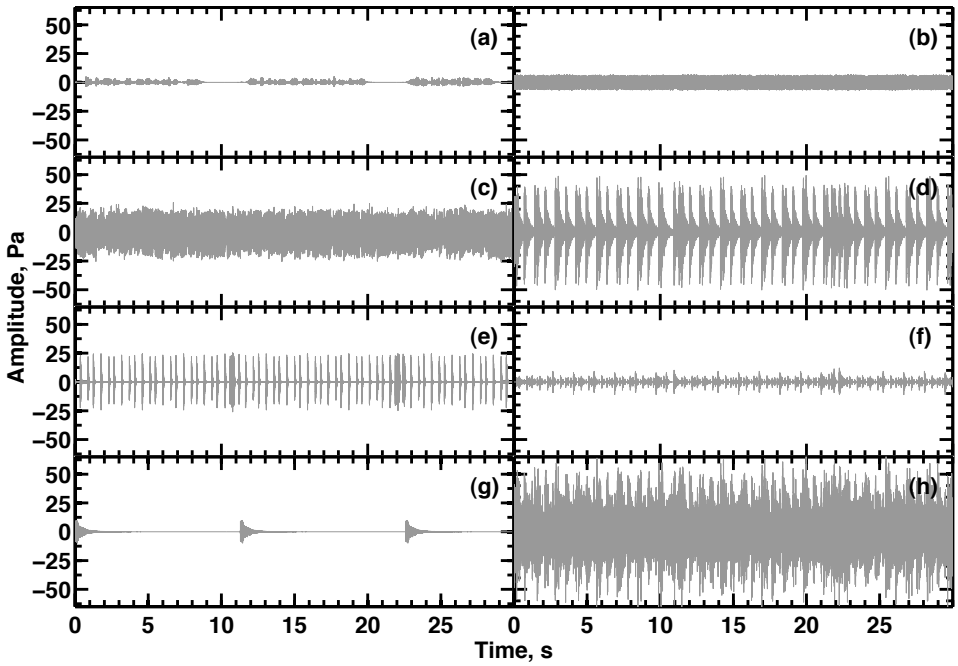


Figure 4.1: The acoustic signals at a reference distance of 1m, plotted in Pascals (Pa), where (a) to (h) correspond to: voice, guitar, bass, kick, snare, hi-hats, cymbal and mix respectively.

(a), which shows that it decreases with increasing venue size.

The optimisation algorithm must be adapted to given control over the overall mix level. A straightforward solution would be to apply an additional constraint to enforce a minimum (or maximum) level. Such a constraint would have a similar form to the feedback constraint, i.e. and additional function that would need to be evaluated for each trial solution. The disadvantage of an explicit constraint is the inevitable increase in optimisation time, particularly when using the genetic algorithm (see Section 3.2 for a discussion of the effects of side constraints on genetic algorithms).

An alternative approach is to adapt the error function used in the first stage of the algorithm, which is repeated here,

$$e_{v_l} = ||MAX\,(0, m_{L_{il}} - m_{R_{il}})\,||. \tag{4.1}$$

The objective of this equation is to make the vocal level sufficiently high in the mix when considering only the direct sound of the other instruments (see Section 3.2.3). The level of direct sound is relatively lower in large venues (due to the increased distances), so a lower absolute vocal level is needed to satisfy Equation 4.1. The other instruments are set relative to the absolute

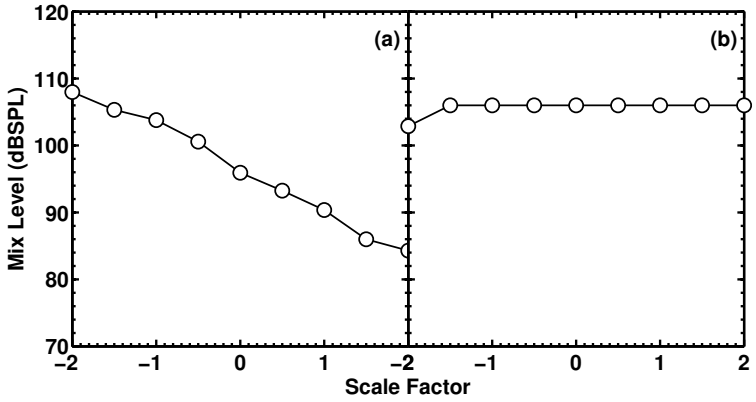


Figure 4.2: The overall RMS mix level at the front centre audience location as a function of venue size; (a) using the original error function to set the vocal gain, Eqn 3.3, (b) using the updated error function, Eqn. 4.2.

vocal level, and if this is lower, then the overall mix level will be lower, as shown in Figure 4.2(a).

Equation 4.1 can be adapted by incorporating an absolute vocal level objective, and substituting the relative mix $m_{L_r}$ with the absolute mix, $m_{L_a}$ (see Equation 2.8). The error can then be evaluated by comparing the actual vocal level to the reference vocal level. If $v_l$ is the element of $m_{L_a}$ corresponding to the vocal level for listener $l$ (element 1 in the our case), then the error function can be rewritten as,

$$e_{v_l} = ||C - v_l||, \tag{4.2}$$

where $C$ defines the reference vocal level, i.e. if $C = 95$ dB SPL, then the error will be zero if the absolute vocal level is 95 dB SPL. Considering only the front centre audience location and incorporating the updated vocal error function, gives the overall mix levels shown in Figure 4.2(b). The level is stable for all venue sizes. For future analysis, Equation 4.2 is used to set the vocal gain, with $C = 95$ dB SPL. The total vocal error, $\varepsilon_{v_T}$, is still evaluated using Equation 3.4.

In some ways, the adapted vocal error function (Eqn. 4.2) is less sophisticated than that used previously (Eqn. 4.1), because it does not take into account the individual requirements of the listeners, or the direct sound levels at different locations. The effect of direct sound is particularly relevant for the drummer, who for any venue size is exposed to the sounds of the drum kit at very high levels. If the target vocal level of 95 dB SPL were delivered to the drummer, the vocals would still be too low when compared to the direct level of the kick drum and cymbal to

achieve the reference mix, as given by Equation 3.1. However, if the original vocal error function were used, then a disproportionate amount of the feedback headroom would be used to raise the drummer's vocal level, particularly for larger venues where the audience demands for absolute vocal level are lower (see Fig. 4.2(a)). Finding the balance between the types of constraints and the requirements of individual listeners is an interesting area for future work. It is likely that more sophisticated requirements should be based on perceptual features, which is particularly true for monitor mixes, where, in the case of the drummer, a more suitable objective would be to make the vocals audible, rather than attempting to recreate a precisely defined mix balance. The simple error function given in Equation 4.2 is used for the remainder of this Chapter, but perceptual features of mixes are explored from Chapter 6 onward.

## 4.3 Venue size

The optimisation algorithm developed in the previous chapter is used to do automatic mixing on venues of different sizes and setups as discussed in Section 4.1. For each venue size the optimisation algorithm is run 10 times, using the two stage optimisation process and a combination of genetic and gradient descent algorithms. Based on the findings in the previous chapter, these algorithms use a genetic algorithm population size of 10, and 10 gradient descent iterations. The best solution within each set is used in the subsequent analysis.

The residual mix error for each venue size is plotted in Figure 4.3. The error is significantly larger for small venues, and for the smallest venue ($d = -2$) is 20 times greater than the baseline venue ($d = 0$). For larger venue sizes the residual mix error decreases gradually, and seems to be approaching a minimum when $d = 1$. Although the general trend is a decrease in error with increasing size, the function plotted in Figure 4.3 is not monotonically decreasing (i.e. for $d = 0.5$ the error is unexpectedly high). This is attributed to the stochastic nature of the optimisation algorithm.

### 4.3.1 Small venues

Table 4.2 contains the mixing desk gain settings and the mix errors for the smallest venue size ($d = -2$). For the FOH mixes, the largest amount of gain is applied to the vocals. Small amounts of gain are applied to the guitar, bass, snare and hi-hats, and negligible gain is applied to the kick and cymbal, showing that the direct sound is significant. The same is true for the monitor mixes,

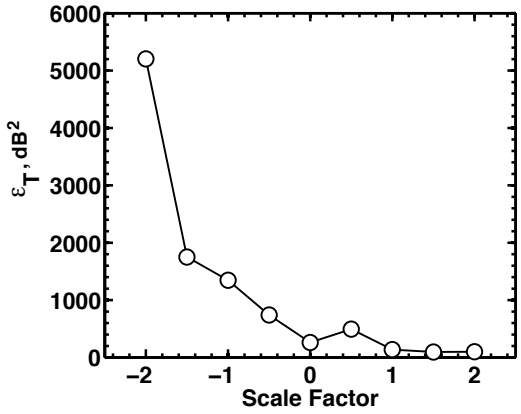


Figure 4.3: The residual error in the mix objective function, plotted as a function of the venue scale factor. The total error is calculated as the sum or squares of the (weighted) error, and hence has units dB$^2$.

with even less gain applied to the non-vocal instruments, so the direct signal path provides almost all of the sound to the monitor mixes.

The errors in the mixes are substantial, and are largely caused by insufficient vocal level. The objective absolute vocal level was 95 dB SPL, and column $v_l$ shows that at most locations it is lower than this, but it cannot be increased further because of the feedback constraint. At the central audience locations there are negative mix errors (vocal too high), which shows that there *is* sufficient vocal level to produce the reference mix. The relatively high vocal level at these locations occurs because they are equidistant from the two FOH loudspeakers, so the sound that reaches them will have combined perfectly in phase at all frequencies. However, the relatively low vocal level at the other audience locations requires a best fit mix, and there are fairly consistent errors across audience locations of $e_l \approx 6dB$. The main cause of the error is the level of the kick drum, which is too high at all locations. For the performers, all non-vocal instruments are too high, and in particular the bass and the components of the drum kit.

The weighting matrix used (see Eqn. 3.2) favoured the audience and the vocalist, and this is reflected in the lower audience errors ($e_l$). As previously discussed, the feedback constraint is a limiting factor, but it is the weighting matrix that determines the attribution of the 'vocal gain resource' amongst the difference mixes (see Sec. 3.3). The larger chunk of this resource is allocated to the FOH loudspeakers, and (effectively) all of that used for the monitor mixes is applied to the guitarist's and bassist's monitor loudspeakers. This is slightly unexpected because the vocalist's needs were favoured over the other performers. However, for such a small venue, the performers are close to one another's monitor loudspeakers, and a better fit to all monitor

| | | $v_l$ | Vocals | Guitar | Bass | Kick | Snare | Hi-Hat | Cymbal | $e_l$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Gain | FOH L | | 6.5 | -4.4 | -6.5 | -23.0 | -3.5 | -4.4 | $-\infty$ | |
| | FOH R | | 6.5 | -4.4 | -6.5 | -23.0 | -3.5 | -4.4 | $-\infty$ | |
| | Mon V | | $-\infty$ | -25.7 | $-\infty$ | $-\infty$ | $-\infty$ | -24.1 | $-\infty$ | |
| | Mon G | | -1.4 | -16.4 | -11.9 | $-\infty$ | -14.6 | $-\infty$ | $-\infty$ | |
| | Mon B | | -6.6 | -21.7 | $-\infty$ | $-\infty$ | $-\infty$ | -24.8 | $-\infty$ | |
| | Mon D | | -376.2 | $-\infty$ | $-\infty$ | $-\infty$ | -4.5 | $-\infty$ | $-\infty$ | |
| Error | Vocalist | 89.5 | 0.0 | 2.3 | 11.8 | 16.7 | 6.9 | 12.0 | 13.7 | 28.3 |
| | Guitarist | 90.5 | 0.0 | 5.3 | 3.3 | 11.9 | 2.0 | 10.3 | 9.0 | 19.3 |
| | Bassist | 87.3 | 0.0 | 0.2 | 11.0 | 11.2 | 5.3 | 8.6 | 14.7 | 23.8 |
| | Drummer | 88.6 | 0.0 | -2.0 | 3.0 | 16.3 | 8.4 | 10.2 | 11.6 | 24.2 |
| | Aud FL | 91.0 | 0.0 | 0.7 | 0.2 | 4.9 | 4.6 | 0.4 | 0.9 | 6.8 |
| | Aud FC | 95.5 | 0.0 | -2.0 | -1.1 | 3.2 | -2.2 | -2.8 | -3.4 | 6.3 |
| | Aud FR | 90.6 | 0.0 | 0.4 | 3.1 | 4.4 | 2.4 | 0.6 | 2.1 | 6.4 |
| | Aud BL | 87.3 | 0.0 | 0.6 | 1.4 | 2.8 | -0.4 | 0.8 | 2.2 | 4.0 |
| | Aud BC | 91.3 | 0.0 | -1.4 | -3.8 | 3.4 | -5.0 | -1.3 | -1.9 | 7.6 |
| | Aud CR | 87.0 | 0.0 | 0.4 | 2.4 | 3.0 | 2.3 | 1.9 | 3.5 | 6.0 |
| | $\varepsilon_T$ | | | | | | | | | 5,204 |

Table 4.2: Control parameters (gains) and residual errors for the optimal solution using a venue size scaled using $d = -2$. The gain section shows the gain applied in dB to each instrument in the indirect path via each loudspeakers. The error section shows the relative error of each instrument, the combined error at each listener location, $e_l$, and the total error for all listener locations combined, $\varepsilon_T$. In addition, the column $v_l$ shows the absolute vocal level for each listener.

mix requirements has been achieved by sharing the loudspeakers. This can be seen in Figure 4.4, which is a plot of the absolute vocal level within the venue (this does not include room reflections). The fact that only two monitor loudspeakers are used suggest that the setup may be improved by deliberately placing the loudspeakers so that the sound they produce is shared amongst all performers, i.e. by angling the two active speakers inward.

The solution to the problem at small venues appears obvious; reduce the level of the direct sound. For amplified instruments such as the guitar and bass this is simple in principle, i.e. turn the volume down, but in practice can be difficult, because musicians are often reluctant to give up control over the level of their instrument, particularly in terms of its contribution to their monitor mix. Encouraging musicians to trust in an automatic mixing system, such as the one proposed, could go some way to alleviating this problem. For loud, acoustic instruments, in particular drums, it is not possible to simply turn the volume down. However, it *is* possible to play instruments like drums with a lower intensity, and the results in this section show that if this is not done, i.e. if a drummer plays at a 'normal' level in a small venue, then it is impossible to produce the reference mix. In small venues the onus is therefore on the musicians to produce sounds at levels that enables the mixing engineer or mixing system to do the mixing task.

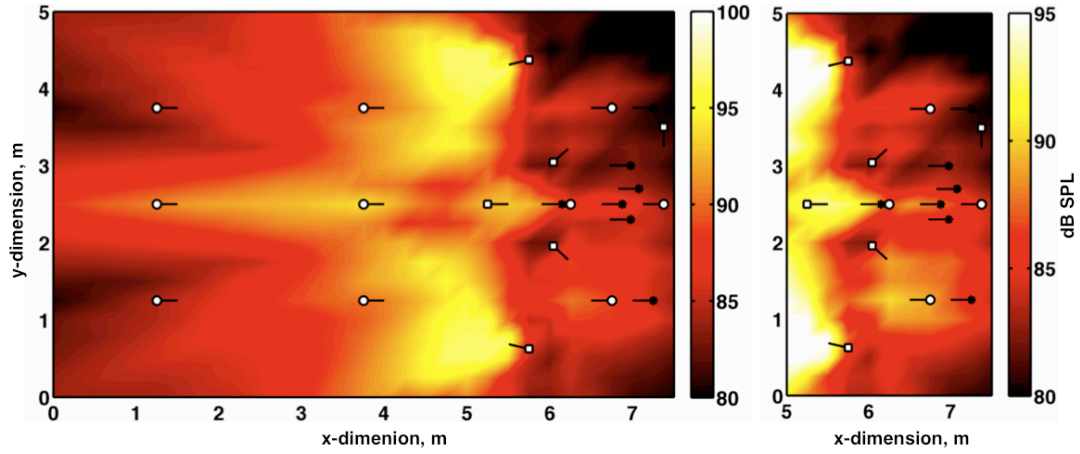


Figure 4.4: The absolute vocal level within for the smallest venue ($d = -2$) in dB SPL. The sound levels have been calculated without including room acoustic effects (note the different scales on the two plots).

### 4.3.2 Large venues

Table 4.3 shows the mixing desk settings and the mix errors for the largest venue size ($d = 2$). As with the small venue size, the largest gains are applied to the vocal signal, however, unlike with the small venue, significant gain is also applied to the non-vocal instruments. This is particularly true for the FOH loudspeakers, showing that the direct sound now has minimal contribution to the FOH mixes. For the monitor mixes the non-vocal gain is lower, showing that direct sound still contributes to the monitor mixes. The most stark difference compared to the smallest venue is the availability of vocal gain for all monitor loudspeakers, and the provision of the reference $v_l$ vocal level of 95 dB SPL (or thereabouts) for all listeners.

The mix errors are far smaller when compared to the small venue. The exception of the drummer, who has too much bass kick drum, hi-hat and cymbal, which as discussed in Section 4.2, is in part due to the use of an absolute reference vocal level. In contrast to the smaller venue, with the exception of the drummer, the monitor mix errors are smaller than those of the audience. This is because the increased distances between the performers de-couple their mixes, so each can be set separately. Although the FOH mixes are now de-coupled from the monitor mixes, the individual audience mixes are still coupled, so a best fit mix is produced. In spite of this the errors in the audience mixes are still very small. The vocal level within the large venue is plotted in Figure 4.5. The distribution of sound within the audience area is similar to that in Figure 4.4, but on stage there are localised hot-spots, showing high vocal level about the performer locations

| | | $\mathbf{v_r}$ | **Vocals** | **Guitar** | **Bass** | **Kick** | **Snare** | **Hi-Hat** | **Cymbal** | $e_l$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Gain | FOH L | | 34.8 | 26.0 | 22.2 | 20.5 | 25.0 | 27.7 | 24.8 | |
| | FOH R | | 34.8 | 26.0 | 22.2 | 20.5 | 25.0 | 27.7 | 24.8 | |
| | Mon V | | 3.0 | -13.0 | $-\infty$ | -12.4 | -6.7 | -13.1 | -11.4 | |
| | Mon G | | 4.2 | -1.2 | -10.8 | -9.2 | -5.1 | -7.3 | -9.7 | |
| | Mon B | | 4.2 | -7.3 | -4.1 | -4.7 | 0.7 | -4.1 | -7.2 | |
| | Mon D | | 3.7 | -5.0 | -6.6 | $-\infty$ | -5.6 | -29.6 | -20.4 | |
| Error | Vocalist | 95.0 | 0.0 | -0.0 | 0.1 | -0.0 | 0.0 | -0.0 | 0.0 | 0.1 |
| | Guitarist | 95.0 | 0.0 | -0.0 | -0.0 | -0.0 | 0.0 | -0.0 | -0.0 | 0.0 |
| | Bassist | 95.0 | 0.0 | 0.0 | -0.0 | -0.0 | -0.0 | -0.0 | -0.0 | 0.0 |
| | Drummer | 94.8 | 0.0 | -0.0 | 0.0 | 7.0 | 0.7 | 3.8 | 5.4 | 9.7 |
| | Aud FL | 95.5 | 0.0 | -0.2 | -0.2 | -0.1 | 0.1 | -0.2 | -0.1 | 0.4 |
| | Aud FC | 98.6 | 0.0 | -0.1 | -0.1 | -0.0 | 0.1 | -0.1 | -0.1 | 0.2 |
| | Aud FR | 95.4 | 0.0 | -0.2 | -0.1 | -0.1 | 0.1 | -0.1 | -0.1 | 0.4 |
| | Aud BL | 92.2 | 0.0 | 0.4 | 0.5 | 0.2 | 0.1 | 0.2 | 0.2 | 0.7 |
| | Aud BC | 95.1 | 0.0 | -0.2 | -0.9 | -0.2 | -0.5 | -0.2 | -0.2 | 1.1 |
| | Aud CR | 92.2 | 0.0 | 0.4 | 0.6 | 0.2 | 0.1 | 0.2 | 0.2 | 0.8 |
| | $\varepsilon_T$ | | | | | | | | | 98.7 |

Table 4.3: Control parameters (gains) and residual errors for the optimial solution using a venue size scaled using $d = 2$. The gain section shows the gain applied to each instrument in the indirect path via each loudspeakers in dB. The error section shows the relative error of each instrument, the combined error at each listener location, $e_R$ and the total error for all listener locations combined, $\varepsilon_T$.

that translates into greater control over the mixes.

## 4.4 Direct sound and mix coupling

Previous chapters and sections have discussed the direct sound and mix coupling in relation to their effect on the mixing process. In this section these phenomena are examined as a function of the venue size.

### 4.4.1 Direct sound

The direct sound is the sound produced by the instruments without any additional amplification. The output sound level of the instruments is fixed regardless of venue size, and its contribution to the overall mix is determined by the distance of the listeners to the sources. Figure 4.6 shows the contribution of the direct sound from each instrument to the mixes at different locations as a function of the venue scale factor. In each case the percentages are shown for: vocal, guitar, bass and the combined drum kit (average of all drum components).They are calculated as linear percentages of the total sound, including both FOH and monitor loudspeakers, i.e. if the absolute levels from direct sound, FOH loudspeakers and monitor loudspeakers are given by D, F and M respectively, in dB SPL, then,

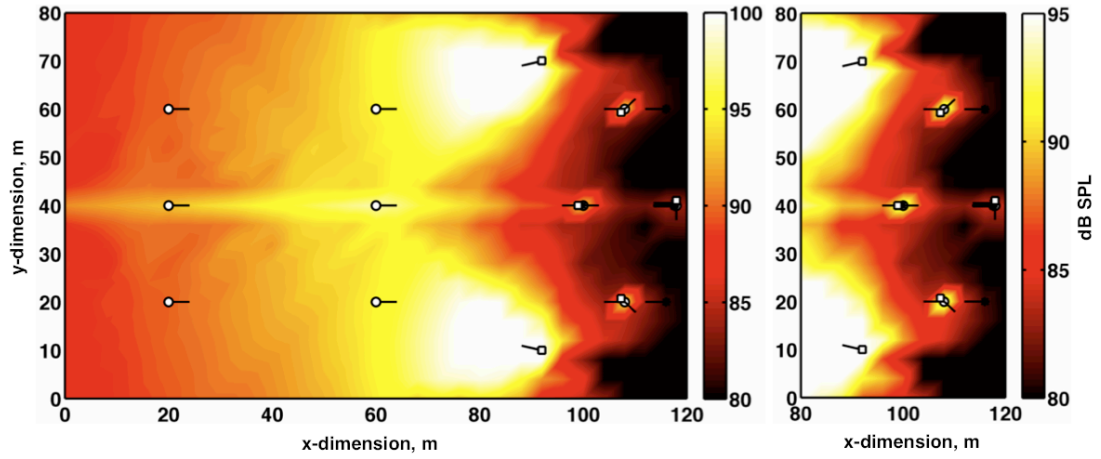


Figure 4.5: The absolute vocal level within for the largest venue ($d = 2$) in dB SPL. The sound levels have been calculated without including room acoustic effects (note the different scales on the two plots).

$$\%D = 100 \times \frac{10^{D/20}}{10^{D/20} + 10^{F/20} + 10^{M/20}}. \tag{4.3}$$

The absolute level of the vocal source is low, so there is minimal contribution of its direct sound to any mix or venue size. Direct sound from all other instruments has a significant contribution to all mixes for small venues, which decreases in larger venues. The exception to this decrease is the drum sound experienced by the drummer, and is caused by his fixed position relative to the drum kit regardless of venue size. The absolute level of the mix is limited by the maximum achievable vocal, so the high levels of direct sound in small venues reduces the control that the engineer has over the mixes, and results in larger mix errors, as demonstrated in Tables 4.2 and 4.3.

Prior research into automatic mixing for live music by Perez-Gonzales and Reiss [2009a,b] focussed on FOH mixes, and made the assumption that the venue is sufficiently large for the direct sound to be ignored. The effect of this assumption can be quantified in terms of the changes it produces in the sound levels. This is shown in Figure 4.7, the markers in which are of the same format as Figure 4.6. It shows that for small venues, large errors in the predicted sound level will occur if the direct sound is ignored. If tolerances were set on the allowable error, Figure 4.7 would show the venue size where existing automatic mixing technology can be used (to set FOH mixes). If the tolerance were set at -2 dB, all venue sizes with errors above the dashed line would be valid for automatic mixing. For the venue setup used here, this gives a minimum

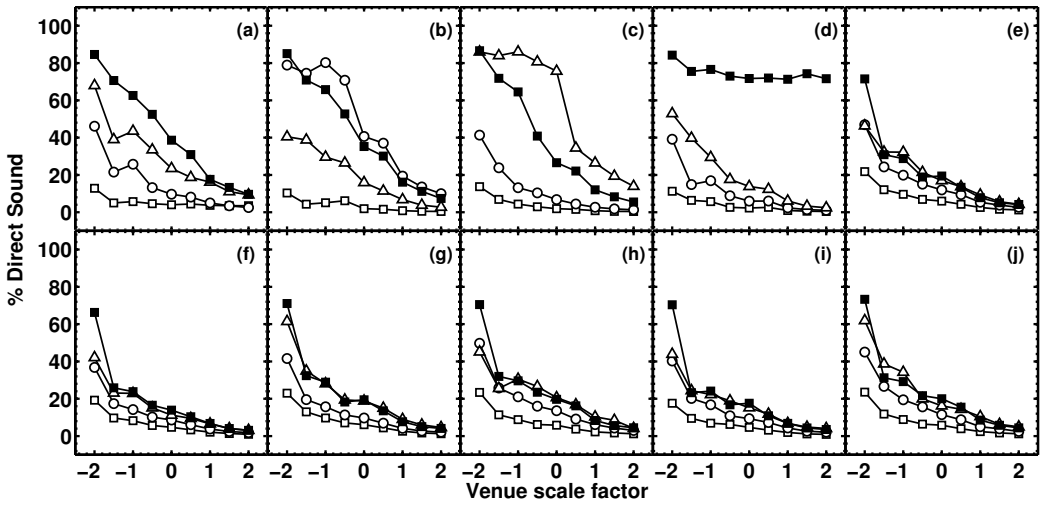


Figure 4.6: The contribution of sound in the direct signal path to the absolute level of each instrument. The white squares, circles and triangles, and the black squares correspond to: voice, guitar, bass, and drums (average across all drum components) respectively. Figs. (a) to (j) are for listeners: vocalist, guitarist, bassist, drummer and audience (FL, FC, FR, BL, BC, BR).

scale factor of -0.6, which is a venue with approximate dimensions of 13m by 20m. Furthermore, Figure 4.7 can be used to classify the 'size' of a performance, inclusive of the absolute levels of the instruments, in terms of these errors, i.e. small performances have errors with magnitudes greater than 2, medium performances between 2 and 0.5 and larger performances less that 0.5. Consultation with the audio industry could lead to a standardised metric describing the size of a live music performance.

### 4.4.2 Mix coupling

In Section 3.3 the coupling between the FOH and monitor mixes was attributed to two factors; (i) the contribution of both sets of loudspeakers to the sound at a given location, and (ii) the sharing of the vocal gain resource, enforced by the feedback constraint. The first of these is examined by plotting the contribution of each set of loudspeakers to the sound levels. Figures 4.8 and 4.9 show the percentage contribution of the FOH and monitor loudspeakers respectively.

Figure 4.8 shows that for each instrument, the percentage of the monitor mixes for guitarist, bassist and drummer, provided by the FOH loudspeakers, is fairly consistent, and is between 20% to 50%. For the vocalist the trend shows a gradual increase in the contribution, up to 80% for the guitar in the largest venue sizes. The monitor mixes are therefore coupled to the FOH mixes. Figure 4.9 shows that for any venue size, a negligible percentage of the audience sound

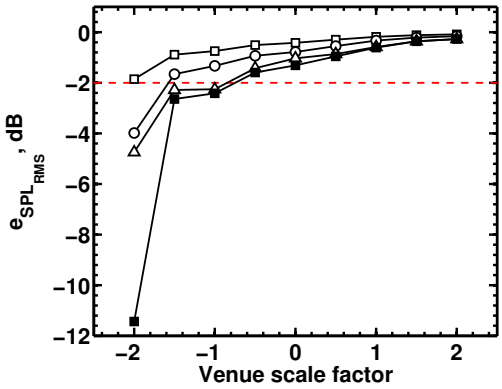


Figure 4.7: The error in the predicted RMS sound level if sound from the direct signal path is ignored. The white squares, circles and triangles, and the black squares correspond to: voice, guitar, bass, and drums (average across all drum components)

comes from the monitor loudspeakers, therefore, in terms of the loudspeaker effect, the FOH mixes are not coupled to the monitor mixes. This means that it is safe to set the FOH mixes without considering what is coming out of the monitor loudspeakers, but that the monitor mixes will change if they are set without considering the sound coming from the FOH loudspeakers.

It was shown in Section 4.3.1 that in small venues there is insufficient headroom to supply the reference absolute vocal level of 95 dB SPL. When there is insufficient headroom, the division of the vocal gain between different mixes is another form of coupling. Figure 4.10 shows the mean residual errors, $e_{V_l}$, calculated in the first stage of the algorithm, for the monitor (circles) and FOH (squares) mixes. For each set, the errors are plotted for mixes produced simultaneously (black lines), and for monitor and FOH mixes produced separately (grey lines). The latter show how well the objective can be met if monitor and FOH mixes are treated as being uncoupled. The monitor mix errors are substantially larger in smaller venues when coupling is included, showing that the coupling effect of the feedback constraint is significant. The effect is less pronounced for the FOH mixes for two reasons. Firstly, the FOH loudspeakers are facing away from the microphone, and are relatively further from it compared to the monitor loudspeakers, and secondly, because the audience requirements are weighted over those of the performers (so they are given a greater share of the vocal gain). For larger venue sizes ($d >= 1.5$), the coupled and uncoupled mix errors are equal for both monitor and FOH mixes, which means that there is no longer any feedback coupling.

The grey curves in Figure 4.10 reveal information about the relationships between the indi-

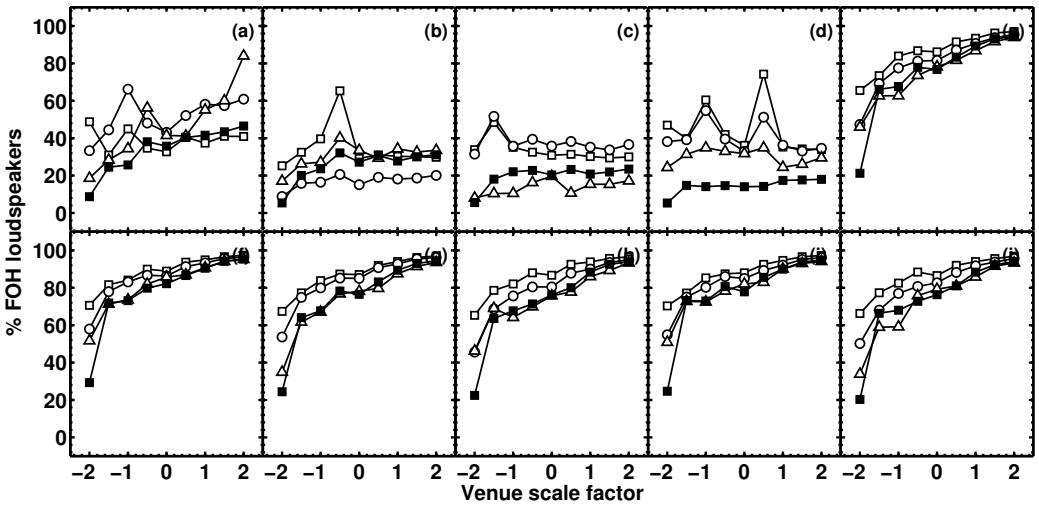


Figure 4.8: The contribution of sound from the FOH loudspeakers to the absolute level of each instrument. The white squares, circles and triangles, and the black squares correspond to: voice, guitar, bass, and drums (average across all drum components) respectively. Figs. (a) to (j) are for listeners: vocalist, guitarist, bassist, drummer and audience (FL, FC, FR, BL, BC, BR).

vidual FOH and monitor mixes. The reference mix is identical for all members of the audience, i.e. we want them all to hear the same mix. They are coupled, because almost all of their vocal sound comes from the FOH loudspeakers, but because they are in different locations within the venue, their absolute vocal levels are different. This is reflected in Figure 4.10, which shows a fixed error in the vocal level of 2 dB for $d \geq -1.5$. The situation is different for the monitor mixes, because each performer has their own designated monitor loudspeaker. When the venue size increases, a point is reached where they are decoupled, enabling the reference vocal level to be provided to all performers, i.e. the error is zero for $d \geq -0.5$. This is shown in Figure 4.5 as the localised hot-spots about the performer locations.

## 4.5 Summary

In this chapter, the automatic mixing system has been applied to virtual live performances in multiple venue sizes. The original error function used to set the vocal level was adapted to provide control over the overall (absolute) mix levels. The mix errors were shown to be larger in small venues, and were attributed to the contribution of the direct sound, the coupling caused by the loudspeaker effect, and by the feedback constraint. In larger venues the contribution of the direct sound was shown to decrease, giving more control over the mixes; and based on the level

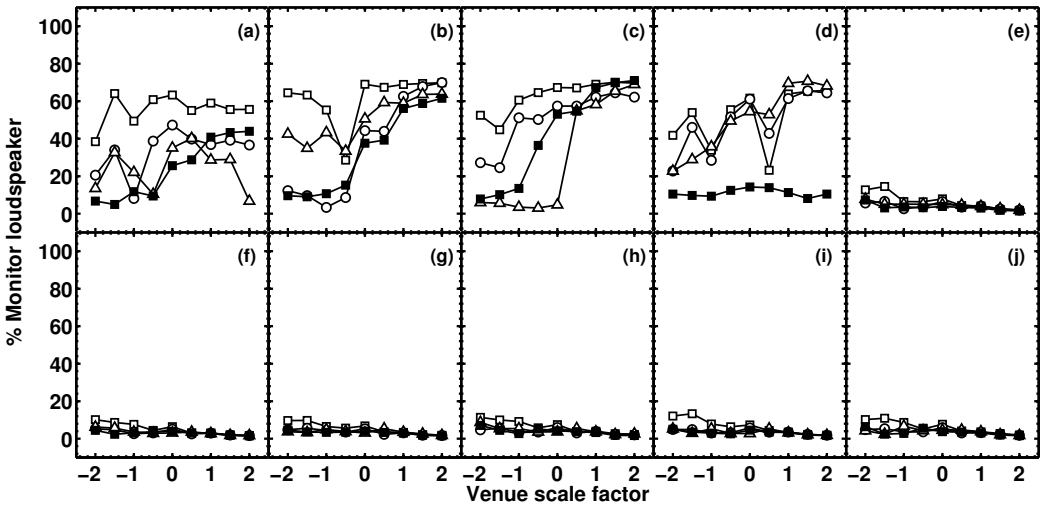


Figure 4.9: The contribution of sound from the monitor loudspeakers to the absolute level of each instrument. The white squares, circles and triangles, and the black squares correspond to: voice, guitar, bass, and drums (average across all drum components) respectively. Figs. (a) to (j) are for listeners: vocalist, guitarist, bassist, drummer and audience (FL, FC, FR, BL, BC, BR).

of the direct sound, a means to describe the 'size' of a performance has been provided. Coupling caused by the loudspeaker effect was shown to be consistent with venue size for the monitor mixes, and negligible for FOH mixes. The coupling caused by the feedback constraint was more restrictive in small venues, particularly on the monitor mixes, but in very larger venues becomes negligible.

The importance of setting suitable vocal levels from which the rest of the mix is based has been highlighted once again. The sensible attribution of the vocal gain resource is critical in providing good mixes. It is somewhat strange that in smaller venues, mixing engineers often set the vocal level last, which subsequently results in a continuous battle to make the voice heard. However, in very small venues, responsibility also lies with the musicians to control their direct sound levels so that it is possible to perform the mixing task.

One of the aims of automatic mixing work is to allow amateur engineers to do live mixing. It is unlikely that a novice will be charged with mixing in a large venue, so in the majority of cases, the effects of the direct sound, mix coupling, and the feedback constraint, will be significant. In order to capture these effects, the models employed must operate on acoustic, as opposed to audio signals, which raises doubts as to the suitability of live automatic mixing systems that use the latter. As well as providing mixing functionality, the system can be employed as an

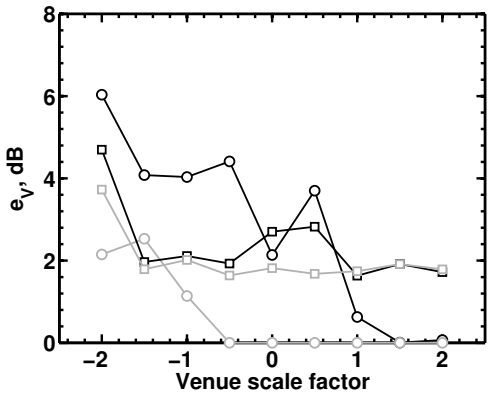


Figure 4.10: The mean residual errors in the first stage of the optimisation for the monitor (circles) and FOH (squares) mixes (calculated using Equation 4.2). For each set, the errors are given when all mixes are considered simultaneously (black lines), and when the monitor and FOH mixes are considered separately (grey lines).

analysis tool. For example, if used in a very small venue, the system could inform the user that certain acoustic signals, e.g. the drums, are too loud to allow the mix objectives to be met. Such knowledge only comes from experience and would be invaluable to a novice mixing engineer.

Some of the plots in this chapter have contained non-monotonic, and slightly erratic curves. In some cases, e.g. Figure 4.3, they are attributed to a local minima, but in others, e.g. Figure 4.9 they are more likely caused by discrete shifts in the structure of the optimal solution. Figure 4.9(a), for example, shows that the percentage of the vocal sound coming from the monitor loudspeaker has a 'zig-zag' form in small venues, and this is caused by shifts in the allocation of vocal gain between loudspeakers, i.e. for $d = -2$ the vocals come from the guitarist's and bassist's monitor loudspeakers, but for $d = -1.5$ the optimal solution is found by sending vocals through the vocalist's loudspeaker. A detailed discussion of these effects is not provided here, but has been mentioned to highlight the difficulty of the task with which an engineer is faced.

In larger venues the contribution of the direct sound becomes negligible because the audience are further from the stage, and can be ignored in automatic mixing applications, as it was by Perez-Gonzales and Reiss [2009a,b]. This also means that the audience is further from the FOH loudspeakers, therefore in order to provide equivalent sound levels, the output level of the loudspeakers must be substantially higher. To provide sufficient sound energy, multiple loudspeakers are used, generally grouped into arrays. The individual components of the arrays will interact, and if poorly managed, can result in vastly different mixes at different audience locations within a venue. If managed well, the mixes will be similar, so the engineer, or automatic

mixing algorithm, can produce a mix at a single position, confident that it will be realised at all audience locations. The control and optimisation of loudspeaker arrays is known as sound system engineering, and is discussed in the following chapter.

# Chapter 5

# Large-Scale Performance and Sound System Engineering

---

In the previous chapter, the effect of venue size on the automatic mixing algorithm was examined. It was shown that the reduced coupling between FOH and monitor mixes, and the reduced contribution of direct sound, make it is easier to deliver the reference mixes. However, in large venues, a substantial amount of sound energy is required, which necessitates the use of multiple loudspeakers. Current best practice groups loudspeakers into arrays, but the interactions between them can be significant and must be controlled. If they can be optimised, such that the transfer functions between the loudspeakers and the audience are uniform, then it is possible for the mixing engineer, or the automatic mixing algorithm, to produce a FOH mix at a single location, confident that the same mix will be experienced by all. This area of work, which is more prominent in industry than academia, is known as sound system engineering. With respect to the work in this thesis, optimising the sound system would be an additional step in automating live production, that would provide a solid base for the subsequent automatic mixing algorithm. Present techniques use both knowledge based and computational optimisation procedures, but at present there is no clearly defined optimisation strategy for controlling loudspeaker arrays. The work in this chapter seeks to provide such a method.

## 5.1 Maximum Sound Pressure Level

Loudspeakers have a limit on the peak level that they can generate. For a given loudspeaker, the peak level is quoted on-axis, at a reference distance (typically 1m, although it can be higher for bass loudspeakers), in dBSPL, and is analogous to the definition of acoustic source signals in

Chapter 2. The data sheets for the front of house loudspeakers approximated in Chapter 3 can be found in Meyersound. The peak level that these loudspeakers can produce is 133 dBSPL, on axis, at a reference distance of 1 m. Figure 5.1 shows the acoustic signals that need to be produced by each loudspeaker in terms of Pascals (Pa), for the largest venue size, and Table 5.1 shows the corresponding rms and peak sound levels in dBSPL for all venue sizes.

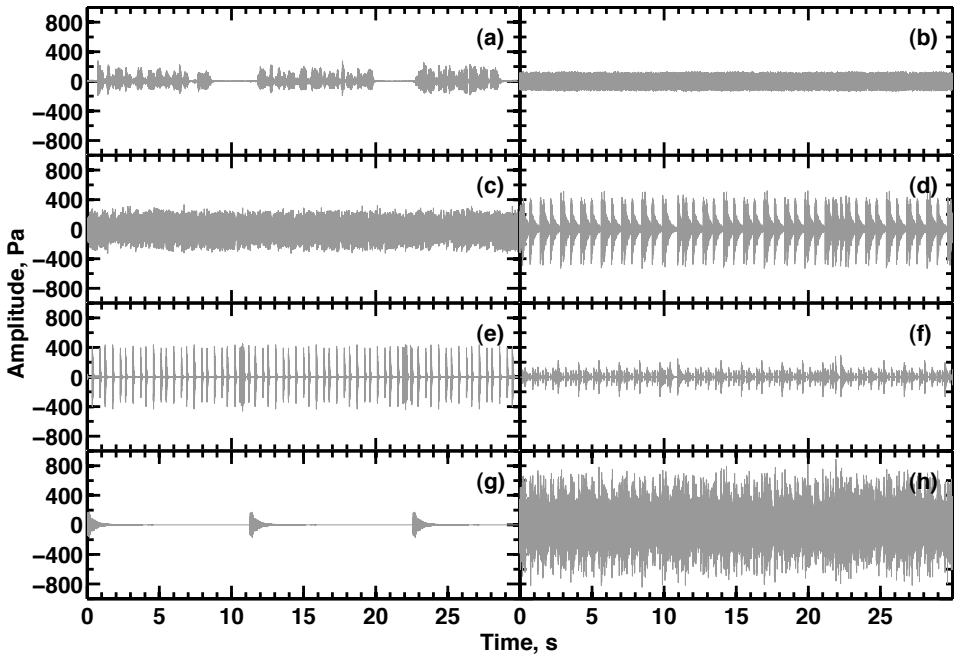


Figure 5.1: The acoustic signals on-axis at a reference distance of 1m from the loudspeaker, for the optimised mix in the largest venue size, (a) to (h) correspond to voice, guitar, bass, kick, snare, hi-hats, cymbal and mix respectively.

Table 5.1 shows that it is impossible to recreate the required acoustic signals with the chosen loudspeakers, for venues with scale factors greater than or equal to 3. By using additional loudspeakers the maximum SPL can be increased, and for each one we add, we get an ideal increase of 6 dB, assuming that the acoustic signals from both loudspeakers are perfectly in phase. Even if this assumption is made, many more loudspeakers are needed to provide the required sound energy, going up to at least four in the largest venue ($133 + (6 \times 4) = 157 > 153.4$). However, this still represents a conservative estimate because; (i) phase relationships between loudspeakers will reduce the perfect summation of 6 dB, (ii) a certain amount of headroom is required to deal with unexpectedly high peaks, and (iii) although the loudspeaker manufactures quote a maximum peak SPL that can be produced, there is no guarantee that a loudspeaker can continually provide this pressure level over an extended period. It is therefore likely that particularly in larger venues,

| | $d$ | Vocals | Guitar | Bass | Kick | Snare | Hi-Hat | Cymbal | Mix |
|---|---|---|---|---|---|---|---|---|---|
| $SPL_{rms}$ | 1 | 96.5 | 97.6 | 106.9 | 92.3 | 100.3 | 87.8 | -$\infty$ | 108.6 |
| | 2 | 103.1 | 106.3 | 115.2 | 115.5 | 106.9 | 97.5 | 92.4 | 119.1 |
| | 3 | 105.5 | 108.7 | 116.3 | 115.8 | 108.7 | 99.6 | 94.7 | 120.0 |
| | 4 | 108.9 | 112.0 | 120.2 | 119.6 | 114.4 | 103.5 | 98.5 | 124.0 |
| | 5 | 110.4 | 113.6 | 122.4 | 121.1 | 114.9 | 105.7 | 100.8 | 125.7 |
| | 6 | 113.9 | 116.8 | 125.0 | 124.5 | 117.2 | 108.8 | 103.9 | 128.7 |
| | 7 | 118.2 | 121.4 | 129.2 | 128.0 | 122.4 | 113.4 | 108.5 | 132.7 |
| | 8 | 122.1 | 125.2 | 132.7 | 132.8 | 125.7 | 117.2 | 112.3 | 136.7 |
| | 9 | 124.8 | 128.0 | 135.6 | 135.8 | 128.8 | 119.9 | 115.1 | 139.6 |
| $SPL_{peak}$ | 1 | 115.0 | 106.4 | 116.3 | 105.4 | 120.8 | 112.8 | -$\infty$ | 123.3 |
| | 2 | 121.6 | 115.2 | 124.6 | 128.6 | 127.4 | 122.5 | 118.2 | 132.9 |
| | 3 | 124.0 | 117.5 | 125.7 | 128.9 | 129.2 | 124.7 | 120.4 | 133.8 |
| | 4 | 127.4 | 120.9 | 129.6 | 132.7 | 134.9 | 128.5 | 124.3 | 137.6 |
| | 5 | 128.9 | 122.5 | 131.8 | 134.2 | 135.4 | 130.7 | 126.6 | 139.4 |
| | 6 | 132.4 | 125.6 | 134.4 | 137.6 | 137.7 | 133.8 | 129.7 | 142.4 |
| | 7 | 136.7 | 130.3 | 138.6 | 141.1 | 142.9 | 138.4 | 134.3 | 146.4 |
| | 8 | 140.6 | 134.1 | 142.1 | 145.9 | 146.2 | 142.2 | 138.1 | 150.6 |
| | 9 | 143.3 | 136.8 | 145.0 | 148.9 | 149.3 | 144.9 | 140.9 | 153.4 |

Table 5.1: The sound pressure levels of the acoustic signals on-axis at a reference distance of 1m from the loudspeaker, for the optimised mix at all venue scale factors, in dBSPL.

many more loudspeakers will be needed. The setup and optimisation of loudspeakers is known as sound system engineering.

## 5.2 Sound system engineering

If a common sound signal is output by multiple acoustic sources, they will interact. This manifests itself as a frequency dependent increase or decrease in amplitude, depending on the phase differences in the signal paths. The set up and optimisation of sound re-enforcement systems is know as sound system engineering, and has evolved to minimise these effects. McCarthy [2007] is a comprehensive text on this subject.

### 5.2.1 Loudspeaker and room equalisation

Early work related to sound system engineering sought to equalise either the loudspeaker or the combined loudspeaker and room response. This was done by modelling and then inverting the relevant transfer function. Greenfield and Hawskford [1991] used FIR and IIR filters in two stages to correct the magnitude and phase response of a loudspeaker. Mourjopoulos [1988, 1994], developed one of the early real-time algorithms to equalise a room response. He modeled the room transfer function using an all-pole model, which he then inverted to give a correction

filter, which was guaranteed to be stable. Issues associated with this approach are discussed by Karjalainen and Mourjopoulos [2005], including the high filter order required when using FIR filters, particularly if a good resolution is required at low frequencies. Ensuring that the filter is stable after inversion is also a problem. Direct approaches overcome the filter stability issues by finding the parameters of the equalising filter directly, rather than modelling and inverting a transfer function, and they often seek to numerically minimise an error function of the form,

$$error = ||H_{EQ}H_RH_L - H_T||, \tag{5.1}$$

where $H_{EQ}$, $H_R$, $H_L$ and $H_T$ are transfer functions corresponding to the equalisation filter, and the room, the loudspeaker and the reference response, which is commonly flat[1]. A problem inherent with standard FIR filters is the linear frequency spacing. Human perception is logarithmic with frequency, and efforts have been made to design equalisation filters with logarithmic frequency spacing, using frequency-warped filters. Warped filters replace the delay elements with all-pass delay elements. Kautz filters are a type of warped filter that have individually tuned all-pass elements that allow any frequency resolution to be realised, i.e. perceptual frequency scales such as the Bark and ERB scales [Smith, 1999]. Further details on Kautz filters, including applications to audio, can be found in Paatero and Karjalainen [2003]. Karjalainen and Paatero [2007] used Kautz filters to form the equalisation transfer function to give more flexibility in low frequency resolution. The poles of the Kautz filters were defined and the coefficients optimised using a least squares method. Ramos and López [2006] incorporated an additional direct step which identified the most perceptually significant peaks in the combined loudspeaker and room response. IIR biquad filters, representative of a parametric equaliser, were assigned to flatten each significant peak of the response. Once all filters were assigned their parameters were optimised using a random search technique, and they highlighted the benefit of IIR filters over FIR filters in terms of reduced filter order. Bank [2008] assigned arbitrary fixed poles to the filters, the values of which were perceptually motivated. The filter coefficients were then optimised numerically. The result is equivalent to the Kautz filter method of Karjalainen and Paatero [2007], but is computationally less expensive.

[1] A flat response means equal sound energy at all frequencies.

### 5.2.2 Sound system optimisation

For live musical performance the situation is more complex because there are many loudspeakers and there are many receiver locations. The objective is to control the interactions between the loudspeakers to give the required response at all locations. This is done by applying different filters to the individual loudspeaker elements, and early work by Meyer [1982, 1984a] led this field of study. Meyer developed a computer simulation of loudspeaker array dispersion that differed from analytical methods because it allowed the dispersion characteristics of real loudspeakers (as opposed to idealised point sources) to be used. This was followed by a digital control system [Meyer, 1984b] that enabled the directivity of the arrays to be controlled and adapted. FIR filters have since been used to control the beam width [van der Werff, 1994] and the directionality [de Vries and van Beuningen, 1994] of loudspeaker arrays.

Line arrays have become the favoured FOH loudspeaker configuration, as highlight by Webb and Baird [2003], because they provide a more consistent frequency response over the audience area. The shape of the arrays have evolved over the years. If arranged in straight lines, the arrays produce a very narrow beam, particularly at high frequencies [Ureda, 2001]. This is advantageous for some applications, but in others where broader coverage in the vertical plane is needed, it can be a disadvantage. In Ureda [2001], "J" and "Spiral" line array shapes are discussed, which give a more even spread of sound energy in the vertical plane.

Advancements in line array design have come about through improved understanding of their modelling. In Staffeldt and Thompson [2004], the modelling errors were evaluated for loudspeakers modelled as point sources; the approach outlined in Meyer [1982, 1984a]. They reported errors of $\pm$ 3 dB up to 8 kHz and $\pm$ 6 dB above 8 kHz, and part of the error was attributed to the errors introduced when measuring the response of a single loudspeaker (low frequencies were not considered because of the difficulties in their measurement at this time). Concerns with the method were mitigated in further modelling studies by Feistel and Ahnert [2007], Feistel et al. [2009], so long as the original loudspeaker measurements were of sufficient accuracy. They concluded that 5 degree spatial resolution was sufficient, and if full impulse responses are stored for each angle, more than adequate frequency resolution can be achieved. Use of impulse responses also preserves phase information, which is critical when combining multiple sources as highlighted by Feistel and Ahnert [2005].

The splay angles of loudspeakers elements within an array were optimised numerically by

Thompson [2006]. The precise optimisation algorithm used was not specified, but this work was an early effort to tackle loudspeaker optimisation computationally. This work has since progressed to include optimisation of FIR filters applied to individual loudspeaker elements in line arrays [Thompson, 2009]. Four objectives were identified: (i) reduction of leakage onto non-audience areas, in particular the stage, which can be viewed in the context of this thesis as a reduction in coupling between monitor and FOH mixes, (ii) controlling the profile of the overall level as a function of distance from the stage, (iii) ensuring smooth frequency response variations, and (iv) enforcing an absolute frequency response profile on the audience area. The authors identified ideal transfer functions that would satisfy the weighted objective function built from these requirements, from which the FIR filter was extracted directly using the inverse FFT. The order of the filters caused unacceptable latency, so they attempted to approximate the filter magnitude response using shorter order filters. This, however, altered the phase relationships, which as discussed previously, are critical when combining loudspeaker responses. The solution was to limit the bandwidth of their method to frequencies above 200 Hz, where the latency of directly inverted transfer functions was deemed acceptable.

The work of Thompson [2006, 2009] was discussed from a practical perspective by Thompson et al. [2011], in relation to software tools. They divided the optimisation into two parts, firstly the layout of the loudspeakers, and secondly the setting of filters. This, it was argued, kept the user involved in the setup process. A brief study that included IIR filters on the bass frequencies was included, which, in order to preserve the phase information of the filter had to be included explicitly in the optimisation algorithm. In Thompson et al. [2011], the authors suggest adaptations to the ideal audience responses that are related to sound perception. They argue that as well as including attenuation due to distance, i.e. a reduction in level when further from the stage, frequency dependent attenuation should also be included. High frequencies attenuate more rapidly that low frequencies, so the response at audience locations further from the stage should include a reduction in high frequencies.

Previous studies into the perceptibility threshold of distortions[2] introduced to the spectral profile of a signal have been undertaken Buckleinm [29], Toole [1986], Toole and Olive [1988]. They found that resonances were perceived more readily than anti-resonances, and that for both resonances and anti-resonances, a larger Q-factor (sharper filter) resulted in distortions being less

[2]The term distortion refers to linear changes in the spectral profiles, and not non-linear distortion.

easily perceived. These results could be incorporated into the objective function, particularly with respect to evaluating differences between loudspeaker array responses.

## 5.3 Robust loudspeaker optimisation algorithm using IIR filters

The objective function for loudspeaker array optimisation is generally a function of the overall magnitude response only, although the relative phase response of individual loudspeaker elements is critical Feistel and Ahnert [2005]. To incorporate the full frequency range without introducing excessive latency, IIR filters must be used, but in order to incorporate their phase response, their parameters must be included in the optimisation algorithm explicitly Thompson et al. [2011]. The disadvantage of this approach is increased complexity in the optimisation algorithm that is yet to be solved. The goal of this chapter is to devise a robust and efficient optimisation strategy that can perform this task.

### 5.3.1 Loudspeaker array modelling

The loudspeaker arrays are modelled as a set of point sources with variable polar frequency responses. Idealised loudspeaker models exist, such as a piston in an infinite baffle Beranek [1954], but in general, loudspeaker manufacturers measure responses of individual loudspeakers in an anechoic chamber, and superimpose these responses in the computational model. Meyersound's tool, Mapp Online Meyersound [2011] can be used to simulate arrays composed of their own branded loudspeakers. Using this tool it was possible to extract polar magnitude responses of Meyersound loudspeakers for 30 frequency points, spaced by 1/3 octave and ranging from 20 Hz to 16 kHz . It was not possible to extract phase responses, so it was assumed that there was zero phase difference at all angles and at all frequencies. In practice this is unrealistic, but the objective here is to develop a robust optimisation algorithm, and realistic phase relationships can be substituted from loudspeaker measurements at a later stage.

At each receiver location the magnitude and phase response from each loudspeaker are calculated, incorporating the effects of distance, angle and the loudspeaker filter parameters. The combined magnitude and phase response are calculated as the summation of phasors, and are given by $R(\omega)$ and $\Phi(\omega)$,

$$R_r(\omega)^2 = \left(\sum_{s=1}^{N_s}(A_s cos(\theta_s)\right)^2 + \left(\sum_{s=1}^{N_s}(A_s sin(\theta_s)\right)^2, \tag{5.2}$$

and

$$\Phi_r(\omega) = arctan\left(\frac{\sum_{s=1}^{N_s} A_s sin(\theta_s)}{\sum_{s=1}^{N_s} A_s cos(\theta_s)}\right), \tag{5.3}$$

where $r$ is the receiver index, $s$ is the loudspeaker index, $A$ is the magnitude and $\theta$ is the phase. Treating the summation in this way is a very efficient way to calculate multiple responses, particularly in array based mathematical software, such as Matlab.

## 5.4 Loudspeaker array optimisation

As stated in Section 5.3, the filter parameters are included explicitly in the optimisation algorithm. The filters used are IIR, parametric equalisers Christensen [2003], and each loudspeaker is assigned an individual filter for every frequency point that is evaluated, i.e. the model used here evaluates the frequency at 30 points, so each loudspeaker has 30 IIR filters, with centre frequencies equal to the analysis points. The gain on the filters is constrained between $\pm$6 dB, and the Q-factor (Q) is constrained between 0.5 (wide response) and 4 (narrow response). The filters at the extremes of the frequency range are shelving filters, and all other filters are peak (positive gain) or notch (negative gain). In addition to the filter parameters, each loudspeaker has a broadband gain parameter between $\pm$6 dB, a delay from 0 to 50 ms, and an angle parameter for orientation in the vertical plane from 0 to -45 degrees. These values were taken from sound system engineering best practice. Rotation in the vertical plane was implemented based on the anchor position of the top loudspeaker in the array (a realistic assumption that is determined by rigging constraints). In addition, the anchor position was included as a parameter, to reflect flexibility in the rigging. Further explanations of the setup are given in Section 5.5.

### 5.4.1 Optimisation strategy

It was shown in Chapter 3 that an optimisation procedure can be made more robust and efficient by dividing it into a number of parts. In Chapter 3 this, (a) allowed the constrained part of the algorithm to be targeted at a smaller parameter set, and (b) reduced linear dependency between the parameters. The optimisation strategy employed here is divided into three sections: setting loudspeaker positions, setting broadband gain and delay, and sequential addition of parametric and all pass filters.

*Loudspeaker position*

Each time the orientation or position of a loudspeaker is changed, the signal paths to each receiver must be recalculated. This is computationally expensive relative to the addition of signal processors, the effects of which can be superimposed on the signal paths. Establishing good orientation and position gives a solid starting point from which the filter parameter sets can be thoroughly searched. Discussions with practicing sound system engineers confirmed this as current best practice, and is demonstrated in Thompson et al. [2011].

*Broadband gain and delay*

The gain parameter will set each loudspeaker to a suitable level from which frequency dependent gain adjustments can be made. The relative delay applied to the loudspeakers affects their phase relationships. Delays are generally used to account for differences in distance between loudspeakers and listeners. A common use is for sound systems that have the main loudspeaker arrays around the stage, but which also have smaller arrays scattered around the audience locations. This will not be an issue in the case studies presented, but the delay is included for completeness.

*Equalisation*

Parametric equalisation filters are used to modify the magnitude and phase response of the loudspeakers. Each loudspeaker has an equalisation filter for every analysis point in the model, i.e. each loudspeaker has 30 equalisation filters that are spaced by 1/3 octave. Attempting to optimise all of these filters in one go would likely be computationally expensive and would suffer from linear dependency issues as discussed in Chapter 3. The strategy employed here is to add and optimise the equalisation filters sequentially. At each stage in the process, a filter of a given frequency is added to each loudspeaker, and the gain and Q parameters are optimised. This is repeated until a filter has been added for each frequency point.

*Objective function*

There a many possible ways to define the objective of loudspeaker array optimisation. In Thompson et al. [2011] the objective is broken down into multiple objectives relating to, amongst other requirements, the target audience response and the amount of leakage. With respect to the audience response, perceptual considerations are suggested, including providing a reduction in sound energy, particularly at high frequencies, when further from the array. Evaluating the ideal objective is a substantial undertaking in its own right, and the focus here is on providing a robust

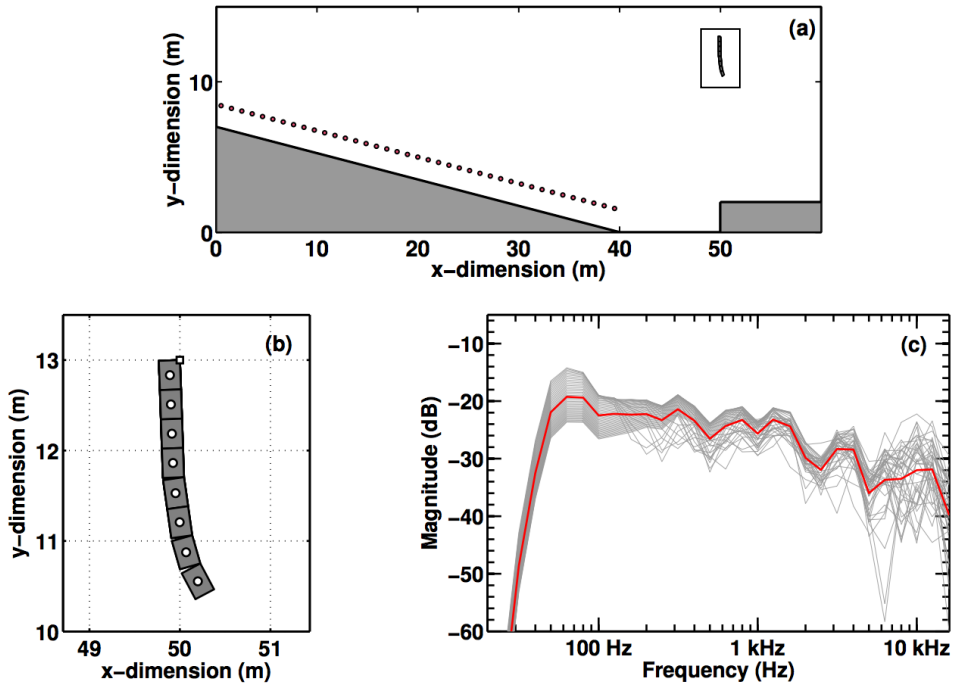


Figure 5.2: The venue layout for case study 1. The receiver locations are identified by circles in (a), and the loudspeaker positions are shown in (b). The grey lines in (c) show the initial response at each receiver location, and the red line is the mean response, and is used as the reference response.

optimisation strategy, regardless of the objective. The simple objective of making the frequency response equal at all audience locations is chosen.

## 5.5 Case study

The case study uses a 2-dimensional loudspeaker array in the vertical plane consisting of 8 loudspeaker elements. The layout of the venue is shown in Figure 5.2(a). The audience are arranged on an incline, and the modelled audience locations are identified by the red circles. The array is anchored at coordinate $(50, 13)$, the first four loudspeakers (from the top) are position with an angle of $-2^o$ to the horizontal, the next two with an angle of $-8^o$, and the final two with angles of $-16^o$ and $-28^o$ respectively, and is plotted in Figure 5.2(b). This setup was chosen to cover all audience locations as evenly as possible, and is in line with best practice McCarthy [2007]. Figure 5.2(c) shows the frequency response at all audience locations with the initial setup. Each grey line represents a different location, and the red line is the mean response. There are substantial differences in the responses between locations, particularly at high frequencies.

### 5.5.1 Objective function

For the purposes of the optimisation, the mean response of the initial setup, i.e. the red line in Figure 5.2(c), is used as the reference response. The selection of a specific response is somewhat arbitrary, because so long as all audience responses are equal, global equalisation can be applied afterward to, for example, give a flat frequency response. By choosing the mean of the initial response as the reference, smaller changes in the control parameters are needed to fulfil the objective.

The objective is to make the response at all audience locations equal to the reference response. However, as it is highly probable that the responses will be imperfect, and that the model of the response will be subject to errors, as highlighted in Staffeldt and Thompson [2004], so the objective is modified to make the responses within a given tolerance of the reference response. The tolerance is set to $\pm$ 3 dB in line with the potential modelling errors suggested in Staffeldt and Thompson [2004]. The error is evaluated using,

$$e = \sum_{r=1}^{N_r} max(|P_{ref} - P_r| - 3, 0), \tag{5.4}$$

where $r$ is the receiver location index, $N_r$ is the number of receivers, $P_{ref}$ is the reference response and $P_r$ is the receiver response.

### 5.5.2 Optimisation parameters

The optimisation process is split into three stages as outlined in Section 5.4.1, and for each stage, suitable parameters are identified that allow a good solution to be obtained in a reasonable amount of time. The parameters explored are: genetic algorithm number of generations and population size, and the number of iterations with the gradient descent search method. All combinations of parameters are trialled with values of: 5, 10, 20 and 40, and for each combination five variants are calculated to account for the quasi-random nature of the genetic algorithm. Based on the findings from previous chapters, the genetic algorithm always precedes the gradient descent method.

*Anchor and angles*

The loudspeaker angles in the vertical plane range from $0^o$ to $-45^o$, and they are constrained such that a loudspeaker must have at least as great an angle as the loudspeaker above. This arrangement is known as 'progressive angles', and is a standard practice in industry, one of the reasons why is the difficulty in rigging if the condition is violated. In addition, the anchor po-

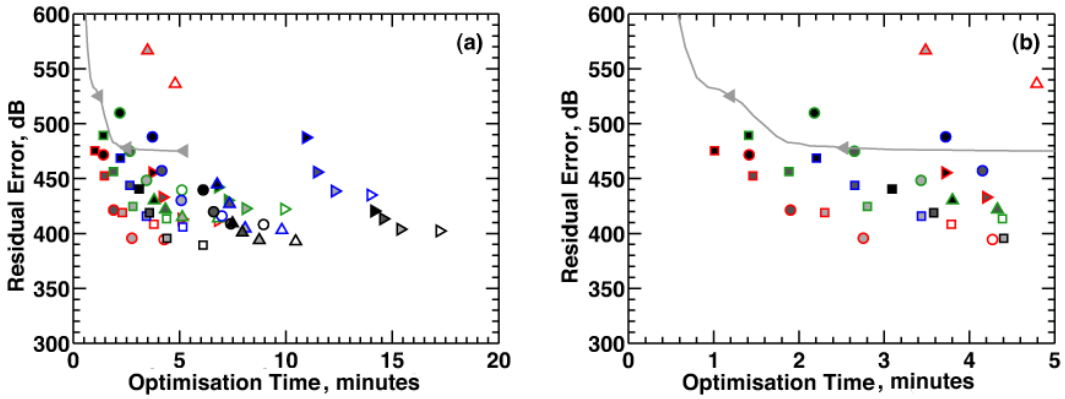


Figure 5.3: The mean residual error in the loudspeaker array error function (Eqn. 5.4), for different optimisation parameter sets, plotted against the optimisation time, using the loudspeaker position. The parameters used are: genetic algorithm generations and population size, and gradient descent iteration number, using values of 5, 10, 20 and 40. In order of increasing parameter value, the colours identify population size: red, green, blue and black; the markers identify generations size: squares, circles, vertical triangles, right-facing triangles; and the shading identifies the number of iterations: black, dark-grey, light-grey, and white. The data in (a) and (b) are identical, but in (b) the x-limits are reduced to improve detail on the shorter time solutions, and the solid grey line is the residual when using the gradient descent search method alone.

sition of the array (the top right corner of the top loudspeaker) is included as a parameter, and represents a venue with variable rigging. The anchor is constrained to within $\pm 1$ m from the initial position, in both $x$ and $y$ dimensions. This leaves 10 parameters in total. Figure 5.3(a) shows the mean residual error plotted against the optimisation time for different sets of optimisation parameters. In this plot, and in all subsequent plots of this form, the parameters are identified as follows, where the order given corresponds to increases in the corresponding parameter. The colours identify population size: red, green, blue and black; the markers identify generations size: squares, circles, vertical triangles, right-facing triangles; and the shading identifies the number of iterations: black, dark-grey, light-grey, and white.

Figure 5.3(a) shows that with the larger parameter values the optimisation time increases up to 18 minutes, and the longest times are seen when the genetic algorithm has 40 generations and a population size of either 20 or 40 (blue and black right facing triangles). Inspection of these two sets of markers shows that the number of gradient descent iterations (the shading) does not have such a substantial cost in terms of the optimisation time. In Thompson et al. [2011] a solution time for optimisation of loudspeaker positions was set at around five minutes, and working upon this basis, around half of the parameter sets take too long. However, there are solutions with low parameter numbers that appear to perform equally well.

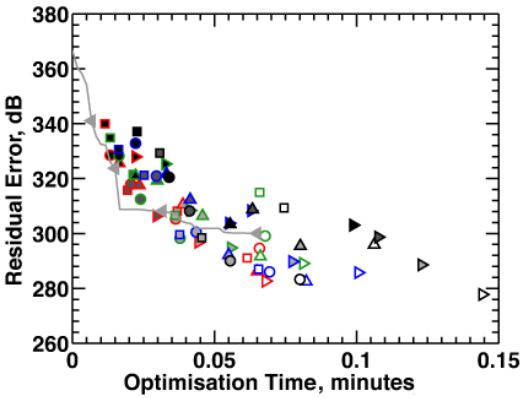


Figure 5.4: The mean residual error in the loudspeaker array error function (Eqn. 5.4), for different optimisation parameter sets, plotted against the optimisation time, using the loudspeaker gain and delay. The parameters used are: genetic algorithm generations and population size, and gradient descent iteration number, using values of 5, 10, 20 and 40. In order of increasing parameter value, the colours identify population size: red, green, blue and black; the markers identify generations size: squares, circles, vertical triangles, right-facing triangles; and the shading identifies the number of iterations: black, dark-grey, light-grey, and white. The solid grey line is the residual when using the gradient descent search method alone.

Figure 5.3(b) shows the same data but zoomed in to include only solution times within 5 minutes. Also shown on this plot as a solid grey line is the residual error when using the gradient descent algorithm only. The grey, left facing triangles are the solutions with 10, 20 and 40 iterations (the datum for 5 iterations is outside of the axes). Using this algorithm alone is relatively fast, and approaches a fairly good solution before becoming stuck in a local minimum. In combining it with the genetic algorithm, we are essentially providing the gradient descent algorithm with an improved starting point, to give it more chance of finding the global minimum. Figure 5.3(b) shows that low valued genetic algorithm parameters, i.e. 5 generations and population sizes of 5 (red squares), still give a good solution if sufficient gradient descent algorithms are used afterward. In order to keep within the 5 minute optimisation time, the solution corresponding to the red circle, filled with light grey, i.e. 10 generations, a population size of 5, and 20 subsequent iterations is chosen. It is worth noting that although the white filled red circle (40 iterations instead of 20) is still within the 5 minute time limit, it does not reduce the residual error.

*Gain and delay settings*

The second stage in the optimisation process sets a broadband gain and delay to each loudspeaker, which are constrained to $\pm 6$ dB and up to 20 ms respectively. The residual is plotted versus op-

| Stage | GA Pop. | GA Gen. | GDS Iter. | Time (s) | Error (dB) |
|---|---|---|---|---|---|
| Start | | | | | 662 |
| Position | 5 | 10 | 20 | 221 | 389 |
| Gain and Delay | 40 | 40 | 40 | 28 | 303 |
| Equalisation | 10 | 10 | 40 | 197 | 141 |

Table 5.2: The parameters used, and residual errors for each stage of the loudspeaker optimisation.

timisation time in Figure 5.5, with the same format as the previous section. Unlike with the loudspeaker position, optimisation of the gain and delay parameters is fast. With the maximum number of parameters it takes 0.15 minutes, i.e. 12 seconds. The trend in the solutions is different compared to the position parameters, in that the residual error continually reduces, albeit gradually, with increased parameter numbers. The gradient descent search (grey line) performs fairly well, but is no quicker than comparably performing combined algorithms. Fortunately this stage in the optimisation algorithm is quick, and so the highest parameter numbers, i.e. the 40 generations, population size of 40, and 40 subsequent iterations (white filled, black right facing triangle), can be used, without much cost in terms of optimisation time.

*Equalisation filters*

The third stage in the optimisation process sets filter gain and Q-factor values of multiple equalisation filters on each loudspeaker. The values of gain and Q are constrained to $\pm 6$ dB, and between 0.5 (smooth) and 4 (sharp) respectively. The filters are added sequentially, in ascending frequency order for each analysis frequency. The model here uses 30 distinct frequencies. The residual is plotted versus the optimisation time in Figure 5.5. Most of the solutions with the low residuals use 40 gradient descent iterations (white filled markers), and there does not appear to be much difference between algorithms that use few or many genetic algorithm parameter values, and this is confirmed by the good performance when the gradient descent algorithm is used alone (grey line). In order to provide a good starting point to the gradient descent algorithm, 10 generations and a population size of 10 are used (green circle), with 40 iterations (white filled).

### 5.5.3 Results

The multistage, loudspeaker array optimisation algorithm, was used to set the array anchor, and the individual loudspeaker positions, gain and delay, and multiple equalisation filters. Table 5.2 summarises the optimisation parameters used for the genetic algorithm (GA) and gradient descent

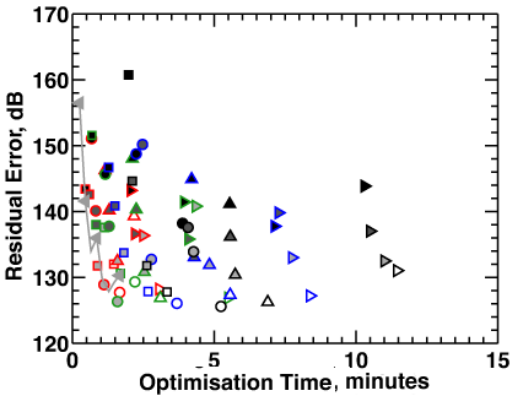


Figure 5.5: The mean residual error in the loudspeaker array error function (Eqn. 5.4), for different optimisation parameter sets, plotted against the optimisation time, using the loudspeaker equalisation. The parameters used are: genetic algorithm generations and population size, and gradient descent iteration number, using values of 5, 10, 20 and 40. In order of increasing parameter value, the colours identify population size: red, green, blue and black; the markers identify generations size: squares, circles, vertical triangles, right-facing triangles; and the shading identifies the number of iterations: black, dark-grey, light-grey, and white. The solid grey line is the residual when using the gradient descent search method alone.

search method (GDS), the residual errors, and the time taken for each stage of the algorithm. The total time for the optimisation was 7 minutes.

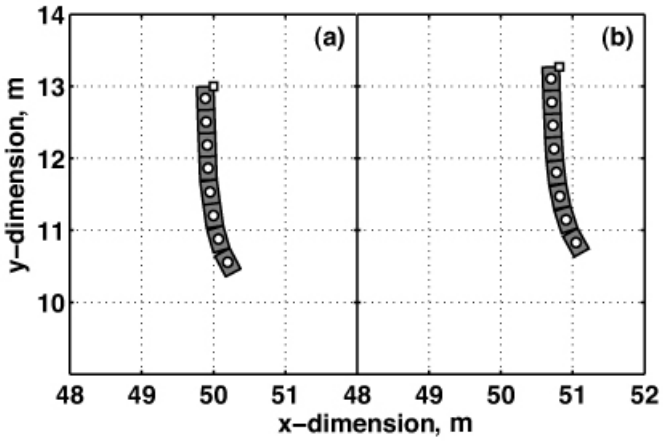


Figure 5.6: The loudspeaker array positions, (a) the initial position, (b) the optimised position. The square in the corner of the top loudspeaker is the anchor point to which the loudspeakers are attached.

The starting and optimised loudspeaker positions are shown in Figures 5.6(a) and (b) respectively. The array has been raised slightly, and angled downward, and the reduction in the error is significant, particularly as the adjustments appear to be relatively minor. The optimised magnitude and phase responses of the loudspeaker filters are shown in Figures 5.7 and 5.8, and include

the effects of gain, delay and equalisation filters. The largest magnitude is 10 dB, and is applied to loudspeaker 1 (top) and 8 (bottom) at low frequencies[3]. The optimisation process described may be used to specify the maximum SPL requirements for each driver in the array, and thus may be used to optimise the monetary cost of the array by using low powered drivers and/or amplifiers where necessary.

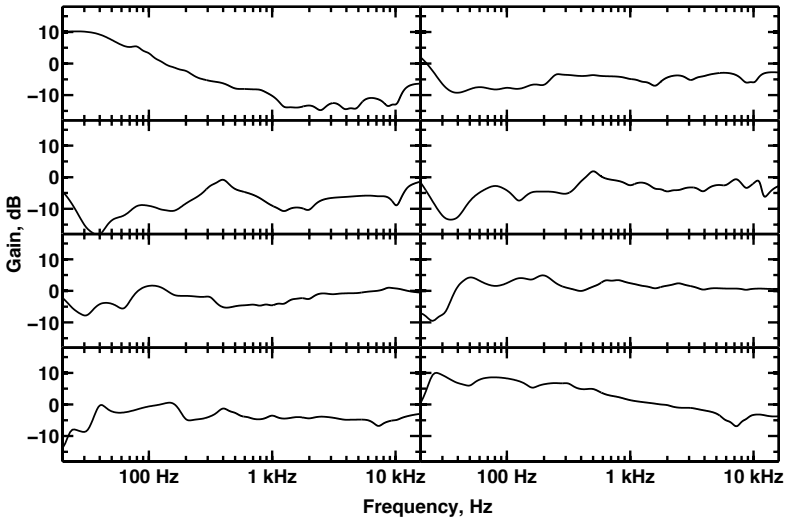


Figure 5.7: The optimised filter magnitude responses, including broadband gain and equalisation filters. (a)-(h) show are for loudspeakers 1 (top) to 8 (bottom), respectively.

The receiver responses at each stage of the optimisation process are plotted in Figure 5.9, and the standard deviation in the error at each frequency point in shown in Figure 5.10. The very low frequency region does not change significantly, so the axes of the response plots have been adjusted to show more detail in the 50 Hz to 16 kHz region. These figures show that loudspeaker position is critical in optimising the response, particularly for frequencies above around 2 kHz. At the higher frequencies, the energy from the individual elements is more focussed, and when combined with the interactions between them, beams of sound energy form, as shown in Figure 5.11, for frequencies of 5 kHz and above. The beams become more narrow for high frequencies, and as a result, neighbouring locations in the audience plane can have vastly different amplitudes. The application of gain and delay provide further control of these interactions, and give an incremental improvement in the frequency response. Where the gain and delay is a relatively blunt tool, the addition of multiple equalisation filters provides fine control over the interactions within each frequency band. Figures 5.9 and 5.10 show a general tightening of the errors at lower

[3]The gain range of individual filters was $\pm 6$ dB, but when combined with the broadband gain, also $\pm 6$ dB, and the neighbouring filters, the net magnitude can exceed this.

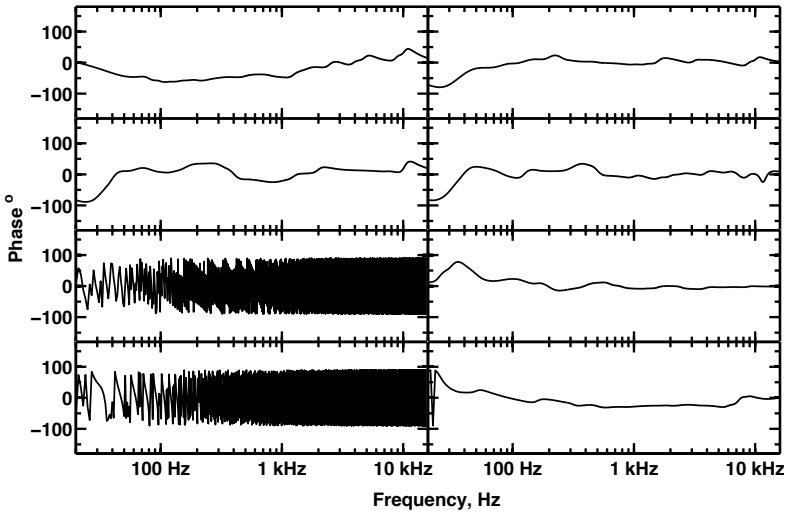


Figure 5.8: The optimised filter phase responses, including delay and equalisation filters. (a)-(h) are for loudspeakers 1 (top) to 8 (bottom), respectively.

frequencies, and targeted response correction at higher frequencies. Figure 5.12 shows that the beams formed at the higher frequencies for the optimised setup are far less pronounced than for the initial setup. The filters have acted to smooth out the high frequency beams, and have given a more even frequency response throughout the venue.

## 5.6 Extensions to arbitrary configurations

The multi-stage algorithm developed provides a robust way to optimise loudspeaker arrays, in terms of the array anchor position, loudspeaker splay angles, individual loudspeaker gain and delay controls, and multiple IIR equalisation filters. The general strategy can be applied to any sound system engineering problem, but the suggested optimisation parameters given in Section 5.5.2 have been derived from a single configuration. Would these numbers change with different configurations?

If the required frequency or spatial resolution were increased the time required to perform each function (error) evaluation would increase linearly, but the number of loudspeakers would be the same, so there would be the same number of control parameters in each stage of the optimisation. With increased frequency resolution it may be necessary to add more equalisation filters, i.e. one per analysis point, but because they are added sequentially, the format of the optimisation stage does not change. Therefore the optimisation parameters stated in Section 5.5.2 can be used for any frequency and spatial resolution, and only need to be reconfigured

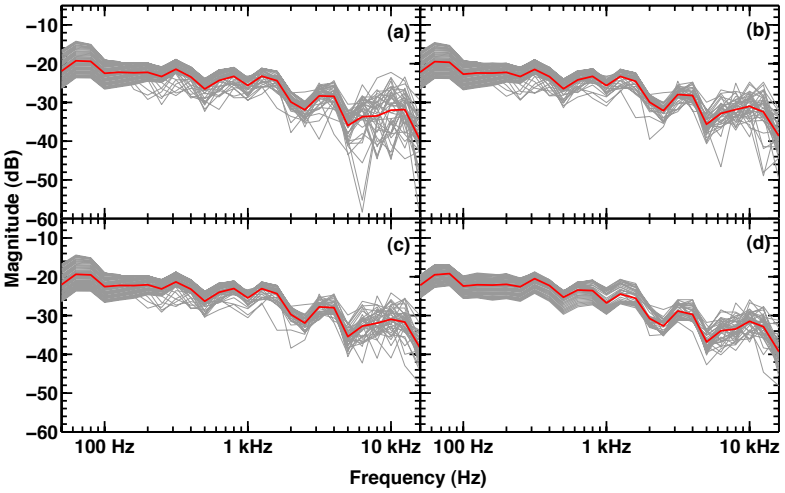


Figure 5.9: The response at each stage of the optimisation, (a) start, (b) positions set, (c) gain and delay set, (d) equalisation filters set. The grey lines are the individual responses, and the red line is the mean starting response, which is used as the reference response.

when there is an increase in the number control parameters, which will occur if there are more loudspeakers.

The receiver response could be improved further by adding more equalisation filters. To do this, the final stage of the algorithm can be repeated an arbitrary number of times, and because they are added sequentially there are no potential problems to the mechanics of the optimisation algorithm. This could be done for all frequency points, or it could be targeted by scanning for problematic areas in the frequency response.

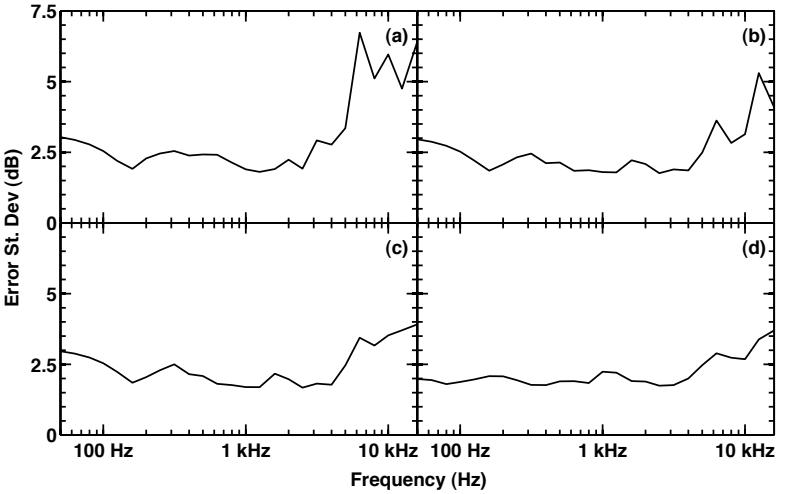


Figure 5.10: The standard deviation in the error for each frequency point at each stage of the optimisation, (a) start, (b) positions set, (c) gain and delay set, (d) equalisation filters set.

## 5.7 Summary

At the beginning of this chapter it was shown that performances in larger venues must use multiple loudspeakers if they are to provide the required sound energy. The literature has been discussed in relation to the techniques used to control the interactions between loudspeakers, with the objective of providing an even frequency response to all audience locations. An algorithm has been developed that splits the optimisation process into three sections, and for each section parameters for the optimisation algorithm have been suggested that give a reliable and robust solution, and which do not require excessive computation time. The algorithm has been demonstrated with a two-dimensional case study, and considerations for extension to arbitrary configurations have been discussed.

Sound system optimisation should be viewed as an additional step in live automatic mixing that precedes the algorithm developed in the previous chapters, providing a solid platform upon which automatic mixing is based. For it to be effectual, the direct sound must have minimal contribution to the FOH mixes, and it is therefore only suitable for larger venues. As mentioned in the summary of the previous chapter, it is unlikely that an amateur would be charged with mixing at a large performance. This work therefore falls into the category of tools to 'assist' in the music production process, and is aimed at professionals. In smaller venues there are fewer FOH loudspeakers (often just two), so the benefits of optimising their interactions would be minimal, although a similar procedure could be used to set their positions, to provide as uniform a coverage as possible over the audience area.

The work in this, and the previous three chapters has examined the practical issues associated with live mixing and has provided models and optimisation strategies that can be used to control them automatically. Using the system that has been developed, an amateur can now approach mixing of live music, in the same way as he would for recorded music. In other words, the two types of mixing have been unified, and future automatic mixing developments can be applied to both. The reference mix that is input to the system is described using objective features extracted from its acoustic signals, but the intention is to extend the work to use features based on their perceived loudness. The loudness of musical sounds is examined in the following chapter.

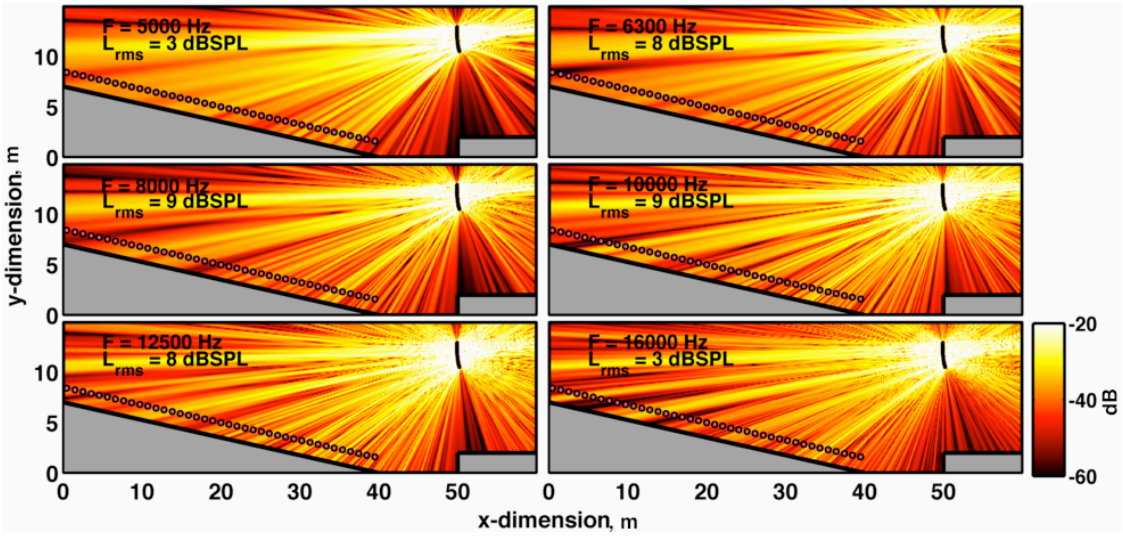


Figure 5.11: Plots showing the distribution of high frequency sound energy within the venue for the initial setup. Each sub-figure is labelled with the frequency it represents, and is quoted with the maximum level of the sound in dBSPL.

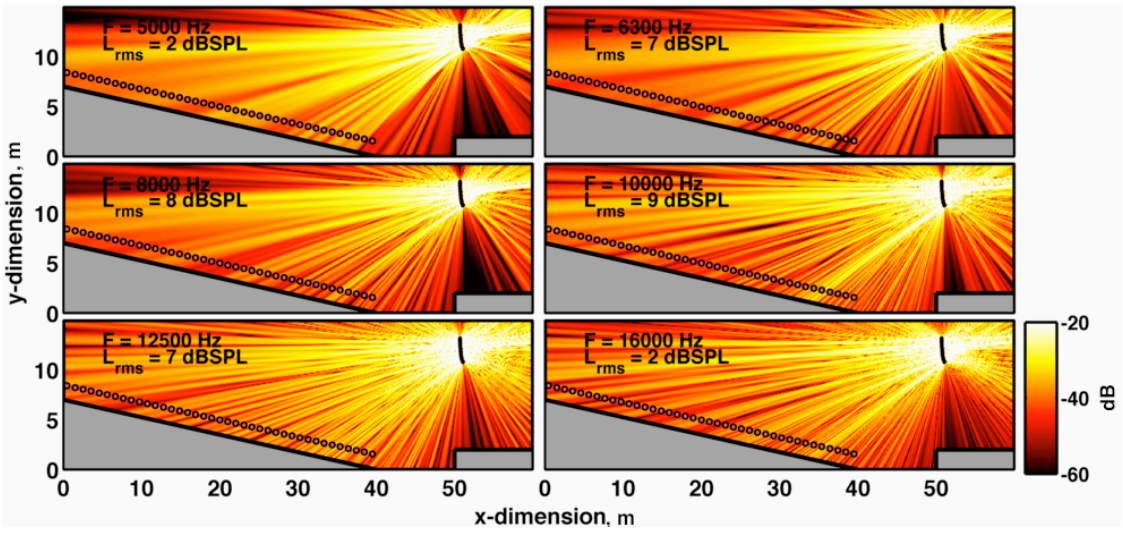


Figure 5.12: Plots showing the distribution of high frequency sound energy within the venue for the optimised setup. Each sub-figure is labelled with the frequeny it represents, and is quoted with the maximum level of the sound in dBSPL.

# Chapter 6

# Estimated Loudness Ratios of Musical Sound-Streams

---

The previous chapters have provided models, and a framework, to do automatic mixing based on objective sound features extracted from the acoustic signals that form the mix. The focus has been on live music performance, where the complexity in the model is greater. The associated issues have been studied in full, and automated solutions for all venue sizes have been provided, and the result is a unification of the approaches to mixing live and recorded music.

An objective of this thesis, as outlined in Section 1.8, is to provide a means to do automatic mixing using perceptual features, i.e. loudness. In order to do this, a loudness feature that is analogous to the relative sound level feature is needed, e.g. the relative loudness or loudness ratios. However, as discussed in Section 1.6.4, although current models are able to estimate loudness time-functions for time-varying sounds, there is no validated means to convert these time-functions into an overall loudness impression, applicable to musical sounds. Furthermore, this raises an important question: is it even possible to describe the relative loudness of two musical sounds using a single ratio? In this chapter, the psychophysical method of *magnitude ratio estimation* is used to answer this question, for sounds played in isolation, and for sounds played in the presence of other masking sounds.

## 6.1 Musical sound streams

The experimental methods used in this chapter require participants to estimate loudness ratios of musical, acoustic signals, which from this point onward are referred to as sounds. Prior work upon which these experiments are based used artificial test tones, and to the best of my knowl-

edge, the work in this thesis is the first attempt to apply this method to musical sounds. As a result, it is useful to make a number of definitions before proceeding.

The stimuli used are generated using pre-recorded audio signals. Each audio signal is a recording of a series of sound events produced by an acoustic source (i.e. an instrument), and is referred to as an *audio-stream*. When the audio-stream is reproduced it is converted into a stream of sound events, i.e. a *sound-stream*. The audio-stream in a digital system is scaled with respect to the maximum allowable sample value in the system (which is referred to as 0 dBFS, where FS stands for full-scale), and the sound-stream is scaled in terms of absolute pressure (dBSPL).

When we listen to a sound-stream we infer some grouping upon the sound events. The term *auditory grouping* originates from Auditory Scene Analysis (ASA) [Bregman, 1978, 1990], and is defined as the process by which concurrent or successive sound events are separated or grouped into streams by the listener. Auditory grouping converts sound-streams into auditory-streams, where each auditory-stream is associated with one or more concurrently active sound sources. When we listen to a stream of sound events produced by a single instrument we tend to group them into a single stream, and when we listen to a mix, we naturally segregate the sounds based on the instruments that produced them. For any psychophysical experiment involving musical sounds, auditory grouping and segregation must be considered.

Stream segregation is known to be promoted by almost any salient physical property of the respective sounds (spectrum, temporal/harmonic relationships) [Moore and Gockel, 2002]. Musical sound-streams produced by different acoustic sources tend to have different properties, so segregation based on the original source is expected. Within the ASA literature, perceptual continuity is a phenomenom that describes the perceived continuation of sounds that are interrupted (masked) by interfering sounds [Haywood and Roberts, 2011]. If sound-streams are presented simultaneously there will be masking interactions (it is even possible that at some points in time one stream may completely mask another), but based on perceptual continuity, it is expected that the perception of the auditory-streams will be preserved. Furthermore, for artificial stimuli (tones) stream segregation is known to build up rapidly for repeated events within the first few seconds of presentation [Anstis and Saida, 1985]. Based on this, it is expected that the grouping of the sound events into separate auditory-streams will also build up over time.

The psychophysical method of magnitude estimation [Stevens, 1975] requires participants to assign numerical estimates of magnitude, or *magnitude ratios*, corresponding to the sensation

of given physical stimuli (see Section 1.4.2 for details). If simultaneous musical sound-streams can be segregated into auditory-streams based on their associated acoustic sources, then it should be possible to perform magnitude estimation upon them. This chapter seeks to explore this hypothesis.

The musical stimuli are defined as musical sound-streams in this thesis, to differentiate them from the artificial test sounds used in prior work. They are described objectively as temporal waveforms, with amplitude in terms of absolute pressure, and can be thought of as the objective correlate to the associated auditory-stream of ASA. This chapter contains two experiments. The first requires the participants to estimate loudness ratios of musical sound-streams, presented independently, and will determine whether overall loudness ratios can be reliably estimated, i.e. it will show whether it is possible to describe the relative loudness of two musical sound-streams using a single ratio. The second experiment is identical to the first, but the sound-streams are presented simultaneously. This will determine whether the streams within a mix can be segregated, and whether their relative perceptual quantities, i.e. loudness, can be estimated.

## 6.2 Experimental method

This section outlines the general methods that are used in both experiments in this chapter, including the procedure, the stimuli and the participants.

### 6.2.1 Procedure

In each test the participant performs loudness ratio estimation on a group of musical sound-streams, using a number scale of their choice (see Section 1.4.2 for details). Sound-streams are either presented separately (referred to as the *solo condition*) or simultaneously with other streams (referred to as the *simultaneous condition*). Details on the mechanism of presentation for the two conditions are discussed in Sections 6.3 and 6.4 respectively. A test interface was built in Max-MSP and is shown in Figure 6.1, which enabled participants to control presentation of the stimuli. Sound-streams were presented using a set of KRK KNS-6400 headphones, connected to a computer using a Tascam US-122L audio interface. The interface and headphones were calibrated such that an audio-stream with a peak at 0 dBFS equated to a sound-stream with a peak level of 112 dBSPL. The tests were conducted in the sound-proofed control room, part of the Media and Arts Technology (MAT) studios, at Queen Mary University of London [Morrell

et al., 2011].

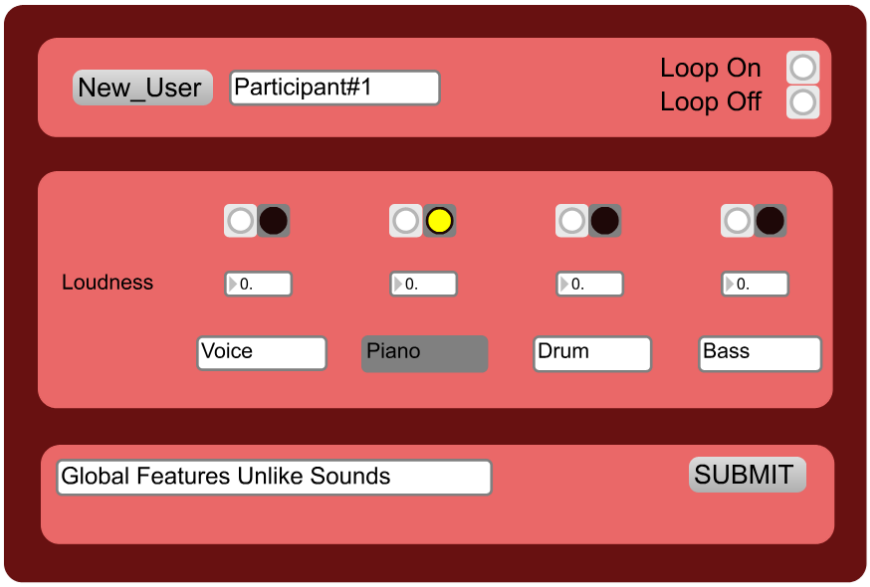


Figure 6.1: The test interface used in the loudness ratio experiment. The loudness ratios are entered in the number boxes above the sound-stream labels. The white button above the number boxes allows the user to play a given sound-stream, the yellow button indicates the currently playing stream, and the grey sound-stream label (piano in this case) identifies the reference stream.

### 6.2.2 Stimuli

The stimuli were produced using four second, pre-recorded, digital audio-streams, where each stream corresponded to a particular acoustic source from: voice, piano, hand-drum and double-bass. The four streams were reproduced simultaneously via headphones, and gain was applied to each audio-stream by an audio engineer to balance the loudness of their associated sound-streams. The temporal waveforms and spectrograms of the balanced audio-streams are plotted in Figure 6.2. The waveforms are plotted relative to the maximum allowable amplitude of the digital system (unity), and the spectrograms are plotted relative to the maximum energy across all audio-streams. The RMS level of the balanced sound-streams were: 87.3, 73.2, 73.6 and 99.5 dBSPL for streams corresponding to voice, piano, hand drum and double bass respectively, and the overall level of the mix was 99.6 dBSPL.

During the experiments the audio-streams were reproduced via headphones with a signal gain applied, between -40 and 0 dB, in 10 dB intervals. The associated sound-streams therefore had an RMS level equal to that set by the mixing engineer when balancing the sound-streams, plus the signal gain. For example, if the audio-stream corresponding to the voice is reproduced with a

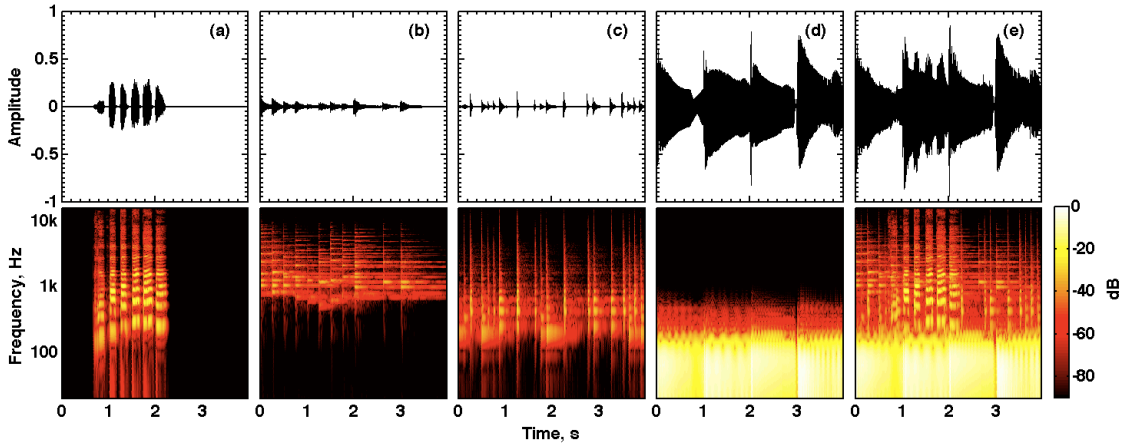


Figure 6.2: Temporal waveforms and spectrograms of the balanced audio-streams, relating to acoustic sources: (a) voice, (b) piano, (c) hand drum and (d) double bass; and (e) is the mix, i.e. the summation of all other audio-streams. The amplitude scaling of the wavefoRMS is relative to the maximum allowable in the digital recording format. The scaling of the spectrograms is relative to the maximum energy across all audio streams.

signal gain of $-10$ dB, the RMS level of the associated sound-stream is $87.3 - 10 = 77.3$ dBSPL.

### 6.2.3 Participants

Twelve adult male participants aged between 21 and 41 took part in each experiment (mean: 28.8, std: 5.5). None reported any hearing impairment, and none had prior experience with magnitude ratio estimation experiments, though all had some experience of amateur audio production.

### 6.2.4 Stream segregation test

Prior to the full experiment, participants were asked to identify by name, the component sound-streams when presented with the mix on a continuous loop, with 0 dB signal gain applied to all audio-streams. Every participant identified the instrument names within 10 seconds, so as expected, the musical sound-streams were segregated into four auditory-streams based on their associated acoustic source.

### 6.2.5 Training

A training phase was used to familiarise the participants with the stimuli and interface, to make them feel comfortable with the process of loudness magnitude estimation, and to give them an impression of the loudness range of the stimuli. They were also asked to estimate loudness ratios for pairs of the same audio-stream, to which different signal gains had been applied. For

example, the audio-stream associated with voice was reproduced with signal gains of -10 dB and -30 dB, giving two voice sound-streams with levels of 77.3 and 57.3 dB SPL respectively. Upon instigation by the participant, the two sound-streams were presented sequentially, separated by a 0.5 second inter-stimuli gap. The participant entered a number for each stimuli that reflected the perceived loudness ratio, and pressed a button to submit the response. This process was repeated for all sound-streams, including the mix, and for all combinations of signal gain. The order of presentation was randomised individually for each participant.

## 6.3 Experiment 1: loudness ratios in solo condition

In the first experiment the sound-streams are compared when presented alternately (the solo condition). The objective of this experiment is to determine whether participants can reliably estimate loudness ratios of musical sound-streams, and will show whether it is possible to describe their relative loudness using a single ratio. In addition, the effect of listen level on the estimated ratios is investigate through adjustments in the signal gain.

### 6.3.1 Procedure

For each test in this experiment the interface loads the four audio-streams and randomly selects a signal gain value from -40 to 0 dB (which it does not reveal to the participant) which it applies to each of the audio-streams. The interface also randomly selects one of the four audio-streams as the reference. A button is available for each stream, which when pressed presents the reference stream, followed by the selected stream, separated by a 0.5 second inter-stimuli gap. For example, if the hand-drum stream is the reference, and the participant selects the voice stream, then the hand-drum stream is presented first, followed by the voice stream. The participant estimates the loudness ratios between the reference stream and all other streams, and is free to repeat the presentation as many times as is necessary within each test, before pressing a button to submit their responses. The test is repeated for each value of signal gain, and with each audio-stream used as the reference. The order of presentation is randomised individually for each participant. This gives two loudness ratios for each pair of stimuli (with both used as the reference), for each value of signal gain, for each participant.

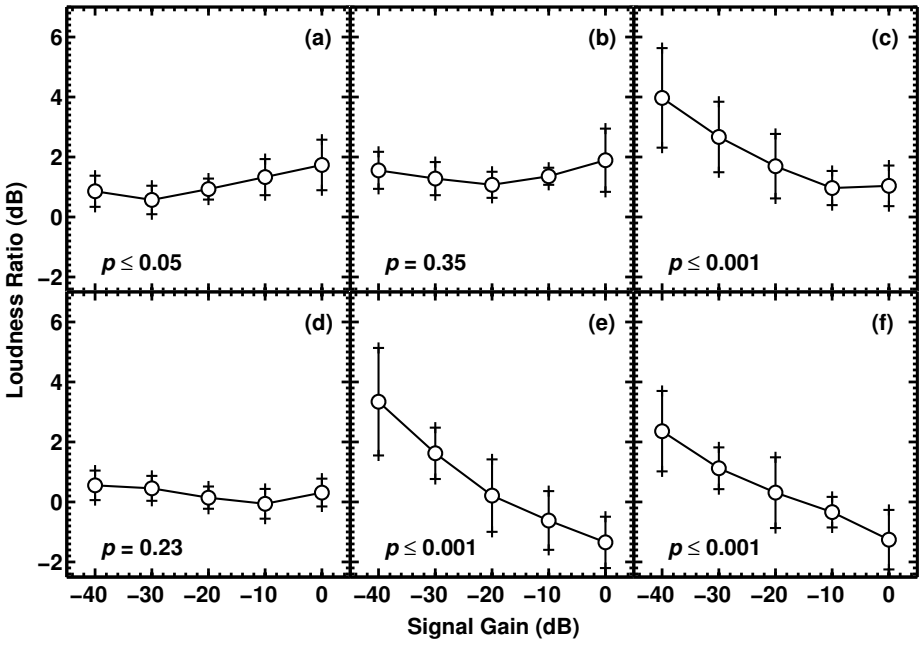


Figure 6.3: The loudness ratio data for the solo condition. Data are the mean and 95% confidence intervals for each combination of stimuli, where (a) to (f) are: voice to piano, voice to hand-drum, voice to double-bass, piano to hand-drum, piano to double-bass and hand-drum to double-bass, plotted as a function of the signal gain. The $p$-values shows the significance of signal gain.

### 6.3.2 Results

The loudness ratio data are converted into decibels, and the mean of the two estimated loudness ratios is taken for each participant, for each signal gain and stimulus pair. The mean and the 95% confidence intervals of the data are plotted in Figure 7.4. For each combination of the stimuli, the estimated loudness ratios are plotted as a function of the signal gain, and Figures 7.4 (a) to (f) are stimuli combinations: voice to piano, voice to hand-drum, voice to double-bass, piano to hand-drum, piano to double-bass and hand-drum to double-bass. The preceding stimuli in each named pair is the numerator when evaluating loudness ratios.

Analysis of variance (ANOVA) tests are performed on the data to determine whether the loudness ratios are significantly affected by signal gain, and therefore by listening level. The output of the ANOVA test is a $p$-value, which is the probability that the mean of the distributions are equal. Its application here compares the distributions in loudness ratios for a given stimuli pair, across all values of signal gain. A low $p$-value, typically less than 0.05, indicates that the means are significantly different. The $p$-values are shown on Figure 7.4. The values obtained

for voice to piano ($p < 0.05$), and for all stimuli combinations containing the double bass ($p < 0.001$), show that the means are significantly different, and therefore that the loudness ratios are dependent on the listening level.

## 6.4 Experiment 2: loudness ratios in simultaneous condition

In the second experiment, the four sound-streams being compared are presented simultaneously (the simultaneous condition). The objective of this experiment was to determine whether participants can segregate simultaneous sound-streams and estimate their loudness ratios.

### 6.4.1 Procedure

For each test in this experiment the interface loads the four audio-streams and randomly selects a signal gain value from -40 to 0 dB (not revealed to the participant) which it applies to each of the audio-streams. The interface also randomly selects one of the four audio-streams as the reference (up until this point, the procedure is consistent with the solo condition). A single 'play' button is available, which when pressed, presents the four streams simultaneously, and an additional control that enables the streams to be played on a continuous loop. The participant estimates the loudness ratios between the reference stream and all other sound-streams, and can spend as much time as required on each set of stimuli before pressing a button to submit their response. The test is repeated for each value of signal gain, with each audio-stream as the reference, and the order of the tests is randomised for each participant. This gives two loudness ratios for each pair of stimuli (with each used as the reference), value of signal gain, and participant.

### 6.4.2 Results

The loudness ratio data are converted into decibels using the same procedure outline in Section 6.3.2, and are shown in Figure 7.5, along with the $p$-values from the ANOVA test. The means are significantly different for all ratios containing the double-bass, and although the $p$-values are above the 0.05 threshold for voice and piano, and voice and hand-drum, they are still low at approximately 0.1, i.e. in these cases, there is a 10% chance that the mean is the same for all values of signal gain.

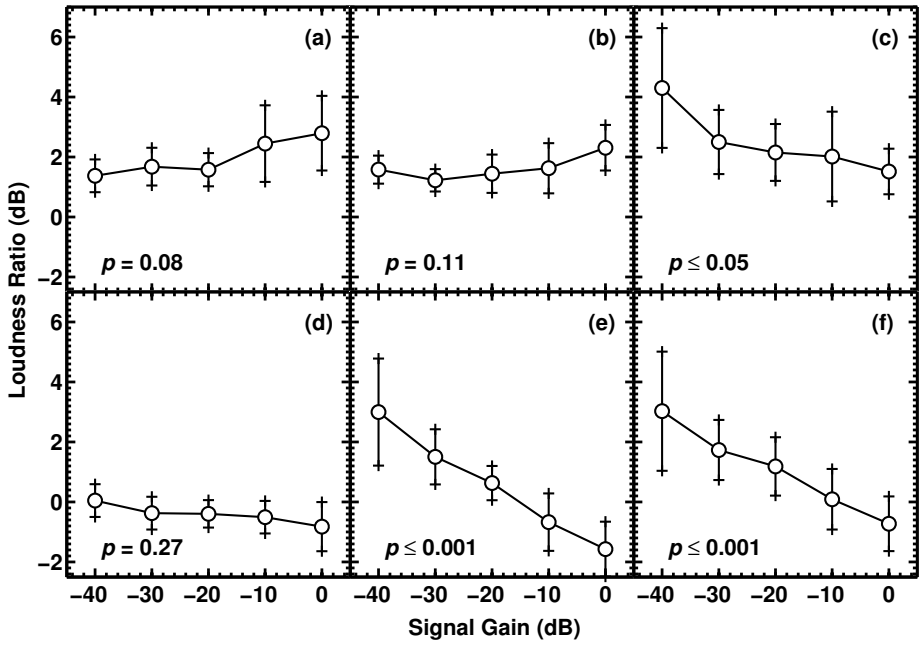


Figure 6.4: The loudness ratio data for the simultaneous condition. Data are the mean and 95% confidence intervals for each combination of stimuli, where (a) to (f) are: voice to piano, voice to hand-drum, voice to double-bass, piano to hand-drum, piano to double-bass and hand-drum to double-bass, plotted as a function of the signal gain. The $p$-values shows the significance of signal gain.

## 6.5 Discussion of experimental data

To the best of the author's knowledge, no prior experiments have applied the psychophysical method of loudness ratio estimation to musical sounds. In fact, before conducting the experiment, there was some doubt as to whether the participants would be able to perform the task reliably. This was borne out by some participants who expressed doubts about the reliability of their responses during the task.

The size of the 95% confidence intervals plotted in Figures 7.4 and 7.5 are shown in Table 6.1 for each stimuli pair. They were evaluated by taking the RMS of the intervals for the five values of signal gain, where the interval size is defined as the upper limit minus the mean. The interval of 0.6 dB for voice versus piano shows that on average (across signal gain), we can be 95% sure that the mean loudness ratio is within 0.6 dB of the estimated mean. Although no statistics are provided, this arguably shows that it *is* possible to describe the relative loudness between two musical sound-streams using a single ratio, and that there is strong agreement between

participants. Although the intervals are larger for stimuli pairs including the double-bass, these loudness ratios are generally larger in magnitude, and therefore a larger variance is expected (this is known as *range bias*, see Stevens [1975] for more details). The intervals for the simultaneous condition are marginally larger than the solo condition, but they are still small. From this it is concluded that participants can reliably segregate simultaneous sound-streams based on their associated sources, and estimate their loudness ratios.

| Condition | V/P | V/D | V/B | P/D | P/B | D/B |
|---|---|---|---|---|---|---|
| Solo | 0.6 | 0.6 | 1.1 | 0.4 | 1.2 | 1.0 |
| Simultaneous | 0.9 | 0.6 | 1.3 | 0.6 | 1.1 | 1.2 |

Table 6.1: The size of the 95% confidence intervals in the experimental data for each stimuli pair. They are evaluated by taking the RMS of the intervals for the five values of signal gain, where the interval size is defined as the interval upper limit, minus the mean.

The $p$-values for the ANOVA tests show that for most stimuli pairs, the mean loudness ratios are significantly different when presented at different listening levels. The automatic mixing algorithm demonstrated in Chapters 3 and 4 used an objective mix description that was independent of listening level. If the reference mix used in this algorithm differed in level compared to the live mix, the perceived loudness ratios may change, even if the objective description is the same. This also applies to pseudo-perceptual features such as the EBU loudness level Union [2011], and the ISO 226 loudness features implemented by Perez-Gonzales and Reiss [2009a,b].

The largest changes in loudness ratios occur with stimuli pairs containing the double-bass. At high listening levels the bass is relatively loud in relation to other sounds, but at low listening levels it is relatively quiet. Figure 6.2(d) shows that the double-bass produces a low frequency sound, so the loss in loudness at low listening levels is consistent with auditory theory, which has shown the absolute threshold of hearing is higher at low frequencies (see Section 1.6.3).

The correlation between the means of the solo and simultaneous data sets is 90% ($p <= 0.001$). In other words, the estimated loudness ratios did not change significantly when the sounds were presented simultaneously. This could be attributable to the fact that the effect of masking was limited. Masking occurs when two sounds overlap in frequency and time. Figure 6.2 shows the spectrograms for the four audio streams. Note that the piano (b), the hand-drum (c), and the double bass (d), all occupy distinct parts of the frequency spectrum, whereas the voice (a) is more broadband, and overlaps with the other sounds in terms of frequency. Some

masking is therefore bound to have occurred, although its effect on the loudness ratios appears to have been limited.

## 6.6 Summary

The psychophysical method of loudness ratio estimation has been used on musical sound-streams in both solo and simultaneous conditions. Participants performed the task reliably and consistently, showing that it is possible to describe the relative loudness of different musical sound-streams using a single loudness ratio. Based on the simultaneous condition, it has also been shown that participants are able to both segregate sounds based on their associated sources, and estimate their perceptual quantities. This suggests that loudness, and partial-loudness ratios, may be good candidate perceptual features to describe a mix.

The loudness ratios changed significantly as a function of the listening level, from which it can be concluded that pseudo-perceptual models of loudness are inadequate for describing mix perception. In addition, it shows that the objective mix description used in Chapters 3 and 4 may break down if the listening level at which the reference mix is set is significantly different from the live performance. The loudness theory discussed in Chapter 1 provides no validated models to estimate the overall loudness impression of musical sounds, hence there is no way to predict loudness ratios. In the next chapter, this limitation is explored, and is overcome, by extending the loudness model of Glasberg and Moore [2002].

# Chapter 7

# Modelling Loudness Ratios of Musical Sound-Streams

---

In the previous chapter, the psychophysical method of loudness ratio estimation was applied to musical sound-streams, in both solo and simultaneous conditions. The estimated loudness ratios were consistent, showing that it is possible to describe the relative loudness of different musical sound-streams using a single loudness ratio, and based on the simultaneous condition, it has also been shown that participants are able to both segregate sounds based on their associated sources, and estimate their perceptual quantities. Loudness, and partial loudness ratios may therefore be used as features to describe a mix, but at present no validated model is available that can predict them. In this chapter, the loudness model for time-varying sounds of Glasberg and Moore [2002] is extended and validated, to predict the overall loudness impression of musical sound-streams, from which loudness ratios can be extracted.

## 7.1 The loudness model

The time-varying loudness model of Glasberg and Moore [2002] is used in this chapter, which is based on the steady sound model of Moore et al. [1997]. The flow chart in Figure 7.1 illustrates the different stages in the model, the input to which is a free-field musical sound-stream. The outer and middle ear transfer functions convert the free-field sound into the pressure at the ear drum, and oval window respectively. Using the 'roex' auditory filter model, and the absolute thresholds of hearing, the excitation pattern across the basilar membrane is evaluated. The excitation pattern is converted into the instantaneous specific loudness (the loudness in each auditory-filter band), which incorporates cochlea compression. The instantaneous loudness is calculated

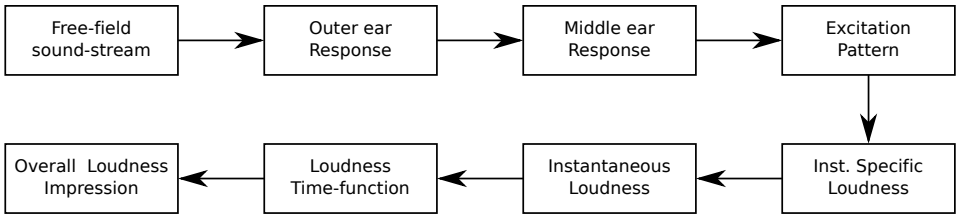


Figure 7.1: A flow chart describing the different stages of the time-varying loudness model of Glasberg and Moore [2002].

by integrating the loudness across all frequency bands, where the prefix 'instantaneous' refers to the fact that the loudness is evaluated for short frames of the sound-stream. The instantaneous loudness is converted into two loudness time-functions: the short-term loudness (STL), and the long-term loudness (LTL), by integrating the instantaneous quantity with time constants that reflect the accumulation of loudness. The STL reflects the momentary loudness of a time-varying sound, and the LTL gives a more long-term loudness impression. The final stage of the flow chart in Figure 7.1, which converts the loudness time-functions into the overall loudness impression, is developed here.

The partial loudness is required for simultaneous sound-streams. A means to estimate this is given for steady-state sounds in Moore et al. [1997], and in Glasberg and Moore [2005], the authors use the same approach for time-varying sounds. The partial loudness is defined by,

$$N'_{total} = N'_{signal} + N'_{noise}, \tag{7.1}$$

where $N'$ is the instantaneous specific loudness, and the subscripts $total$, $signal$, and $noise$, represent the total loudness, the loudness contribution of the signal of interest, and the loudness contribution of the noise (masker), respectively, i.e. $N'_{signal}$ is the partial loudness of a signal in the presence of noise. This equation is rearranged to,

$$N'_{signal} = N'_{total} - N'_{noise}. \tag{7.2}$$

Equation 7.2 is adapted to apply to simultaneous musical sound-streams. If $s(i)$ is the sound-stream generated by instrument $i$, and there are $n$ instruments, then the instantaneous specific *partial* loudness of instrument $i$ is given by,

$$N'(s_i) = N'\left(\sum_{x=1}^{n} s_x\right) - N'\left(\sum_{x=1}^{n} s_{x(x\neq i)}\right). \tag{7.3}$$

In other words, the partial loudness of a sound stream within a set of streams is its contribution to the loudness of the set as a whole. For the purposes of this thesis, the noise signal is taken to be the summation of the 'other' sound-streams in the set. Once the partial instantaneous specific loudness has been evaluated, the standard procedure is followed to give partial STL and LTL.

The loudness time-functions, STL (blue lines), and LTL (red lines), of the four sound-streams used in the previous chapter are shown in Figure 7.2 for each value of signal gain (top to bottom is 0 dB to -40 dB). The STL captures the momentary changes in loudness, whilst the LTL introduces substantial smoothing effects to give a longer term loudness impression. The experiments in the previous chapter showed that the relative loudness between two musical sound-streams can be described by a single loudness ratio. At present, the loudness model provides time-functions, so the objective here is to convert these into a single quantity that describes the overall loudness impression of a sound, from which loudness ratios can be evaluated. This is indicated as the final stage in the flow chart in Figure 7.1.

## 7.2 Modelled loudness ratios

Features of the loudness time-functions are required that describe the overall loudness impression of the sound-streams, for both solo and simultaneous conditions. Candidate features include the mean and peak of the loudness time-functions. The peak of the loudness time-function was suggested by Zwicker [1977] for speech, and the mean LTL was suggested by Glasberg and Moore [2002]. Furthermore, the loudness matching tasks discussed in Section 1.6.4 showed that at equal loudness, steady-state and amplitude modulated sounds had different objective correlates, i.e. rms and peak intensity respectively. The findings of these experiments were attributed to experimental bias by Moore et al. [1998], but they suggested that for sounds with large amplitude modulations, cochlea compression would likely introduce differences in the objective correlates to loudness, when compared to their steady-state counterparts. The sounds used in this thesis are musical, and have different, and complex, dynamic characteristics. It is therefore expected that the feature of the loudness time-function that best describes the overall loudness may be signal dependent, i.e. it could be the peak for some sounds, and the mean for others, or some combination of the two.

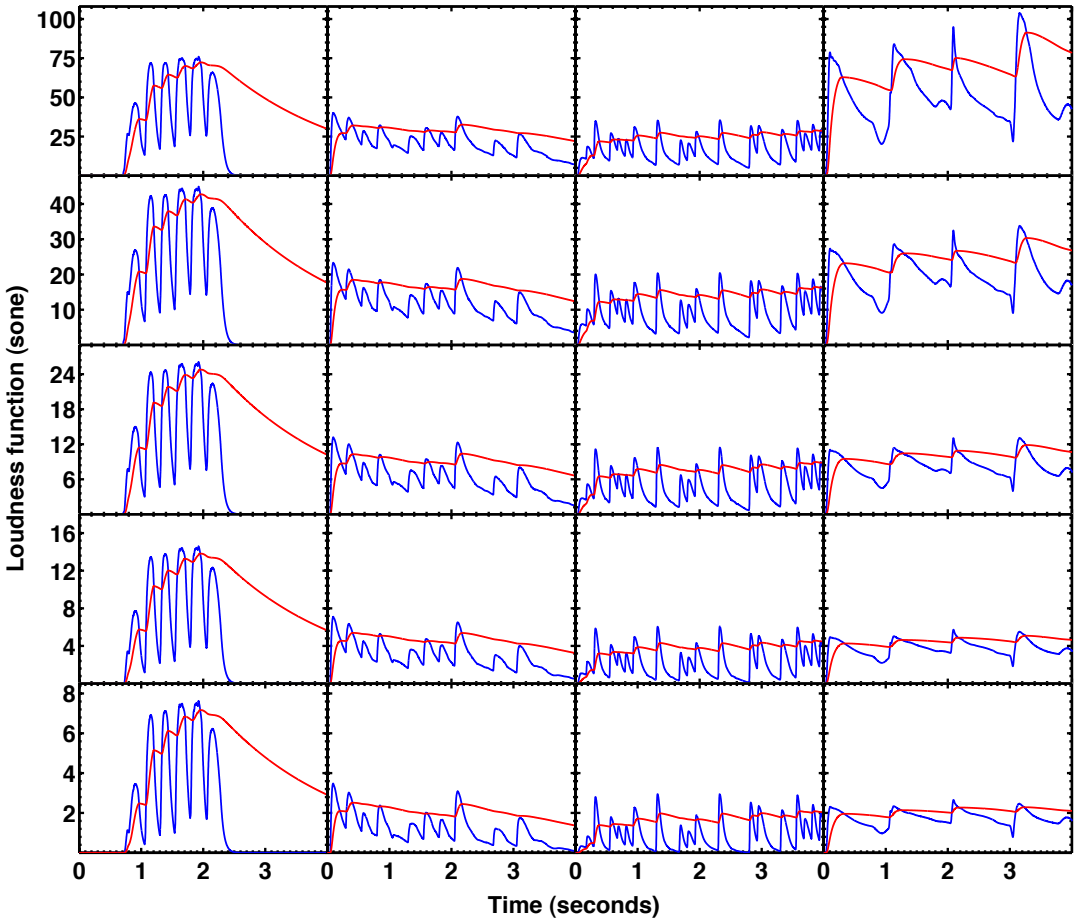


Figure 7.2: The time-varying loudness functions of the sound-streams for each signal gain value. The red line is LTL and the blue line is STL. From top to bottom to signal gain goes from 0 to -40 dB in 10 dB increments, and from left to right the loudness functions correspond to the voice, piano, hand-drum and double-bass respectively.

### 7.2.1 Signal specific bias coefficient

To account for potential differences in the loudness time-function features, the parameter $\alpha$ is introduced. It adds a stream specific bias that weights the contribution of the mean and the peak of the loudness time-function to the overall loudness impression as follows,

$$|L(t)| = \alpha L_{\mu}(t) + (1-\alpha)L_{p}(t), \quad where \quad 0 \leq \alpha \leq 1, \tag{7.4}$$

where $L(t)$ is the loudness time-function (i.e. STL or LTL), the subscripts $\mu$ and $p$ represent the mean and peak values of this function, and $\alpha$ is the stream-specific bias coefficient. Based on Equation 7.4, when $\alpha = 1$ the mean of the loudness time-function is used, and when $\alpha = 0$ the peak is used. The overall loudness impression of a sound-stream $i$ is denoted by $l_i$, and the loudness ratio of stream $i$ relative to stream $j$, denoted by $r_{ij}$ is,

$$r_{mod_{ij}} = 10log_{10}\left(\frac{l_i}{l_j}\right), \tag{7.5}$$

where the subscript *mod* identifies that the loudness ratio has been modelled.

### 7.2.2 Optimal $\alpha$ values

Values of $\alpha$ for each sound stream that best fit the experimental data in both solo and simultaneous conditions are found numerically using a gradient descent optimisation algorithm. The error function evaluates the difference between the experimental and modelled loudness ratios, i.e.,

$$e_{sol} = ||r_{sol_{exp}} - r_{sol_{mod}}||, \tag{7.6}$$

$$e_{sim} = ||r_{sim_{exp}} - r_{sim_{mod}}||, \tag{7.7}$$

$$e_{comb} = ||e_{sol}, e_{sim}||. \tag{7.8}$$

where $r_{exp}$ and $r_{mod}$ are vectors containing the experimental and modelled loudness ratios respectively, and the subscripts $e_{sol}$, $e_{sim}$ and $e_{comb}$ refer to errors in the solo, simultaneous and combined data respectively. The $\alpha$ values were found by minimising $e_{comb}$, and the values are shown in Table 7.1, for both the STL and LTL time-functions.

| Instrument | $\alpha_{STL_{opt}}$ | $\alpha_{LTL_{opt}}$ |
|---|---|---|
| Voice | 0.49 | 0.79 |
| Piano | 0.21 | 0.0 |
| Hand-drum | 0.0 | 0.0 |
| Double-bass | 1.0 | 1.0 |

Table 7.1: The optimised values of the $\alpha$ coefficient for each sound-stream, when STL and LTL are used as the loudness time-function.

### 7.2.3 Modelling results

The correlation between the experimental and modelled loudness ratios, as well as the error data, are used to evaluate the quality of the fit. Figure 7.3 contains plots of the experimental loudness ratios versus the modelled loudness ratios, where (a) to (f) use features: mean STL, peak STL, STL with optimised $\alpha$, mean LTL, peak LTL, and LTL with optimised $\alpha$ respectively. The red and blue markers identify the solo and simultaneous conditions respectively. Markers lying on

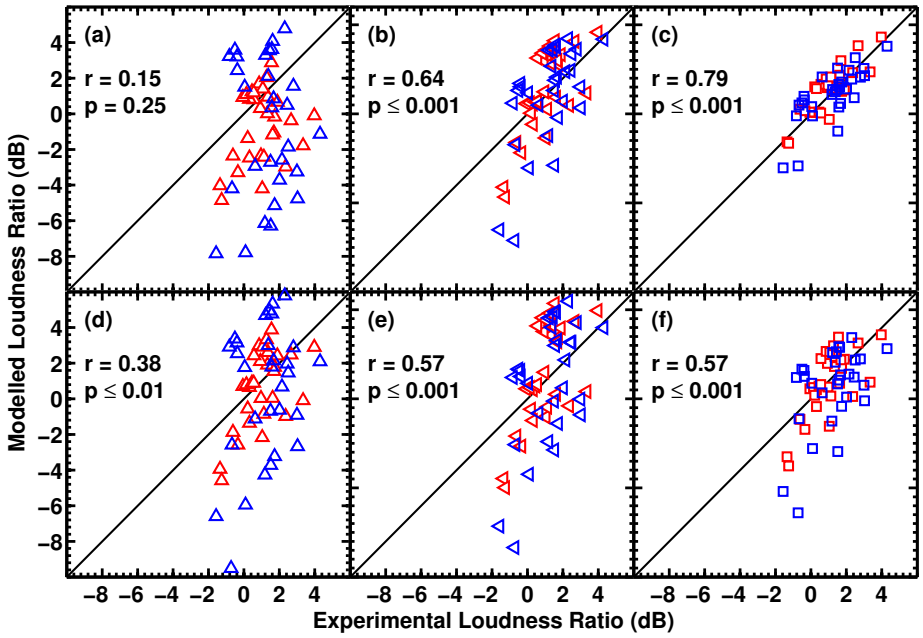


Figure 7.3: The experimental loudness ratios plotted against the modelled ratios, where (a)-(f) are: mean STL, peak STL, STL with optimised $\alpha$, mean LTL, peak LTL and LTL with optimised $\alpha$. The red and blue markers are the solo and simultaneous data respectively. The black line is $x = y$, i.e. markers on this line show an exact match between experimental and modelled loudness ratios.

the solid black line show an exact fit between the model and experiment. The $r$-value shows the correlation, and the $p$-value shows the probability that the correlation could be caused by a random distribution. Figure 7.3(c), which uses STL with optimised $\alpha$, has the highest correlation and a very low $p$-value.

The correlation statistics and the error metrics are shown in Table 7.2 for the solo and simultaneous data, and for all data combined. $\alpha$ values of 1 and 0, correspond to mean and peak of the loudness time-function, and *opt* refers to the optimised values shown in Table 7.1. The lowest error, as well as the highest correlation are found using the optimised $\alpha$ values and the STL. It is interesting to note that the worst performing features are the mean STL and LTL, which is surprising, because the mean LTL is said to give a good impression of the overall loudness in Glasberg and Moore [2002]. From this it can be concluded that the process of estimating the overall loudness of a musical sound-stream is not a simple averaging of the loudness time-function.

| | | Solo | | | Simultaneous | | | Combined | | |
|---|---|---|---|---|---|---|---|---|---|---|
| L(t) | $\alpha$ | r | p | $e_{sol}$ | r | p | $e_{sim}$ | r | p | $e_{comb}$ |
| STL | 1 | 0.27 | 0.15 | 14 | 0.12 | 0.53 | 26 | 0.15 | 0.25 | 30 |
| | 0 | 0.72 | ($\leq$ 0.001) | 8.4 | 0.6 | ($\leq$ 0.001) | 12 | 0.64 | ($\leq$ 0.001) | 14 |
| | opt | 0.86 | ($\leq$ 0.001) | 3.5 | 0.75 | ($\leq$ 0.001) | 4 | 0.79 | ($\leq$ 0.001) | 5.3 |
| LTL | 1 | 0.58 | ($\leq$ 0.001) | 9.1 | 0.3 | 0.11 | 20 | 0.38 | ($\leq$ 0.01) | 22 |
| | 0 | 0.64 | ($\leq$ 0.001) | 12 | 0.53 | ($\leq$ 0.01) | 16 | 0.57 | ($\leq$ 0.001) | 20 |
| | opt | 0.7 | ($\leq$ 0.001) | 6.2 | 0.5 | ($\leq$ 0.01) | 11 | 0.57 | ($\leq$ 0.001) | 13 |

Table 7.2: Correlation coefficients and error between experimental data and model (Eq. 7.4). Values for *opt* denotes the optimized values of $\alpha$ shown in Table 7.1.

### 7.2.4 Predicted loudness ratios

Figures 7.4 and 7.5 contain plots of the experimental and modelled loudness ratios for solo and simultaneous conditions respectively. In each plot, the dashed black line is the mean of the experimental data, and the dashed red lines are the upper and lower 95% confidence intervals. Also shown on each subplot is the rms error, and rms confidence interval for each stimuli pair, the latter of which are taken from Table 6.1. In eight of the twelve pairs the error is smaller than the confidence interval, and even for those where the error is larger, the maximum difference is 0.4 dB. This is a substantial improvement when compared to using the mean or peak of the loudness function, which although not studied in detail for music signals, may be considered to represent the current state of the art.

## 7.3 Dynamic sound-stream bias

Individual values of $\alpha$ have been identified for each sound-stream, which weight the contribution of the mean and peak of the loudness time-function when estimating the overall loudness. The best fit was found when the coefficients are applied to the short-term loudness, suggesting that the long-term loudness fails to capture the dynamic information in the loudness time-function. This is evident in Figure 7.2, which shows the smoothing effect of the LTL calculation, and the loss detail in the transients in the piano and drum sounds. However, the worst performing model used the mean STL, so even though the STL captures the transients of the sounds, simply averaging it does not give the overall loudness impression.

Table 7.1 shows that the values of $\alpha$ (when using STL) are very different across the sound-streams. The value for the hand-drum is 0, i.e. the overall loudness is the peak STL, and for the double-bass it is 1, i.e. the overall loudness is the mean STL. Prior work showed that the loudness of modulated and steady-state sounds had different objective correlates, i.e. rms or

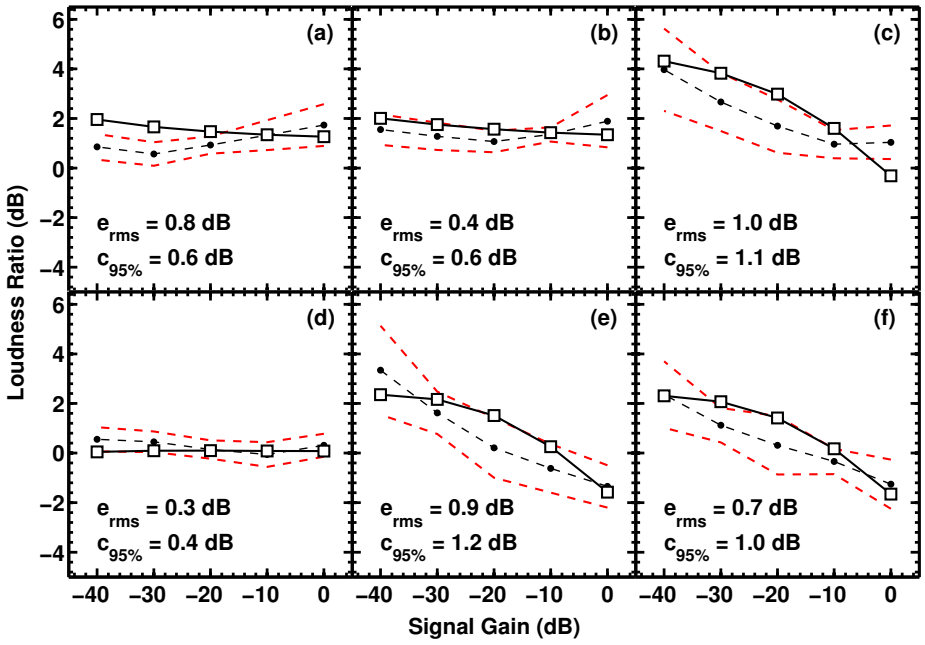


Figure 7.4: The experimental and modelled loudness ratio data for the solo condition. The experimental mean and 95% confidence intervals are shown by the dashed lines for each combination of stimuli, where (a) to (f) are: voice to piano, voice to hand-drum, voice to double-bass, piano to hand-drum, piano to double-bass and hand-drum to double-bass, plotted as a function of the signal gain. The square markers show the modelled loudness ratios using the STL and the optimised $\alpha$ values.

peak intensity. A measure of the amount of modulation in a sound is its crest factor, which is the ratio between its peak and rms amplitude, expressed in decibels. Figure 7.6 is a plot of the alpha value of each sound-stream versus its crest-factor, which clearly shows a strong negative correlation ($p < 0.05$). A linear curve fit yields the equation,

$$\alpha = 1.81 - 0.08c, \tag{7.9}$$

where $c$ is the crest factor of the sound stream in dB. This equation states that for steady-state sounds, e.g. the double-bass, $\alpha = 1$, so the loudness impression is equal to the mean of the loudness time-function, but for very transient sounds, e.g. the hand-drum, $\alpha = 0$, so the loudness impression is equal to the peak of the loudness time-function. Sounds that lie somewhere on the continuum from steady-state to transient, will use a combination of the mean and peak of the loudness time-function. $\alpha$ is termed the *dynamic sound-stream bias* (DSSB) coefficient, and can

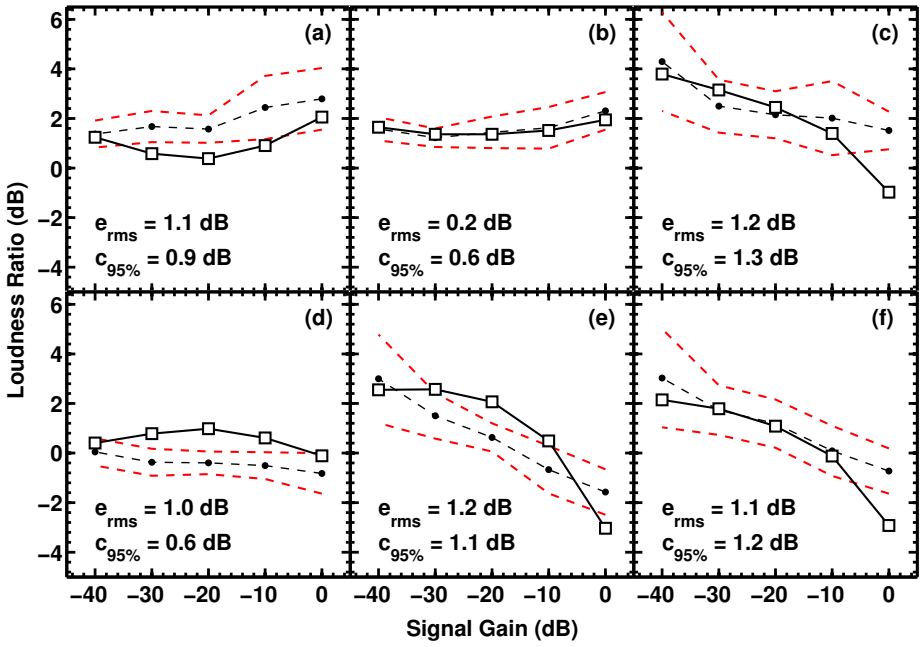


Figure 7.5: The experimental and modelled loudness ratio data for the simultaneous condition. The experimental mean and 95% confidence intervals are shown by the dashed lines for each combination of stimuli, where (a) to (f) are: voice to piano, voice to hand-drum, voice to double-bass, piano to hand-drum, piano to double-bass and hand-drum to double-bass, plotted as a function of the signal gain. The square markers show the modelled loudness ratios using the STL and the optimised $\alpha$ values.

be predicted by substituting the crest-factor of a sound-stream into Equation 7.9.

## 7.4 Summary

The experimental loudness ratios detailed in the previous chapter have been modelled, the basis of which is the time-varying loudness model of Glasberg and Moore [2002], which produces short-term, and long-term, loudness time-functions. Candidate features of these functions have been extracted, and have been used to model the experimental data. A dynamic sound-stream bias (DSSB) coefficient has been introduced and optimised for each sound-stream, which provides a substantial improvement in the loudness ratio predictions, compared to simple mean or peak features. The DSSB coefficient has been shown to be strongly correlated with the crest factor of the sound-streams, enabling it to be estimated for any new stream.

The loudness ratio feature proposed in this chapter is a significant advancement of existing automatic mixing work [Barchiesi and Reiss, 2010, Perez-Gonzales and Reiss, 2009a,b, Terrell

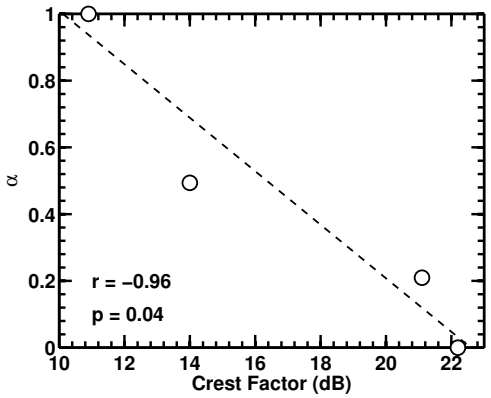


Figure 7.6: The experimental and modelled loudness ratio data for the solo condition. The experimental mean and 95% confidence intervals are shown by the dashed lines for each combination of stimuli, where (a) to (f) are: voice to piano, voice to hand-drum, voice to double-bass, piano to hand-drum, piano to double-bass and hand-drum to double-bass, plotted as a function of the signal gain. The square markers show the modelled loudness ratios using the STL and the optimised $\alpha$ values.

and Reiss, 2009]. It utilises a full auditory (loudness) model, which operates on sound-streams as opposed to audio-streams, and has been validated for musical sound-streams. It incorporates level effects, which can be significant, and perceptual biases based on sound-stream dynamic characteristics; and when applied to simultaneous streams it can account for masking interactions. In the next chapter, the loudness ratios within a mix are expressed as the loudness balance, which is the first validated feature vector, which describes the loudness relationships in a mix as a whole.

The relative sound level of the acoustic signals in a mix, used as the mix feature in Chapters 3 and 4, can be replaced with the loudness balance. This will enable the perceptual features of a reference mix to be automatically recreated at any live performance (within the bounds imposed in a live environment). As discussed in Chapter 1, the objectives of this thesis are not limited to live performance, but include the development of a framework to do automatic mixing of all forms. In the next chapter, as well as considering the implications of loudness balance on the live music work, general applications in automatic mixing are considered.

# Chapter 8

# Loudness Balance: A Perceptual Mix Descriptor

---

In the previous chapter, the loudness model of Glasberg and Moore [2002] was extended to predict the overall loudness impression of musical sound-streams, in both solo and simultaneous conditions, from which loudness ratios were evaluated. This introduced the dynamic sound-stream bias (DSSB) coefficient, which weights the contribution of the mean and peak loudness time-function to the overall impression of loudness, and which is a function of the sound-stream crest factor. In this chapter, the loudness ratios of all sound-streams within a mix are grouped into a mix feature vector, termed the loudness balance. The loudness balance feature is demonstrated in a number of music production applications, including: a live automatic mixing case study, an overview of a perceptual mixing system, a method to extract best-practice features for fully automatic mixing applications, and the definition of a perceptual audio format.

## 8.1 Loudness balance

A common theme in all automatic mixing work is the need to describe the mix. In most cases this is done using objective features, i.e. a spectral histogram [Kolasinski, 2008] or relative intensity levels [Terrell and Reiss, 2009], but pseudo-perceptual features have been incorporated [Perez-Gonzales and Reiss, 2007, 2010], by implementing a loudness model based on the ISO 226 loudness curves [for Standardization, 2003]. However, none of these approaches incorporate the effects of listening conditions or masking interactions, and so do not accurately model mix perception. The loudness balance, which *does* incorporate these factors, is defined in this section, and is arguably the first truly perceptual mix feature.

The partial loudness of each musical sound-stream within a mix is predicted using Equation 7.4, where $\alpha$ is calculated using Equation 7.9. The loudness time-function, $L(t)$, is the short-term partial loudness output by the model of Glasberg and Moore [2002], which incorporates the partial adaptation given by Equation 7.3. The loudness of each stream is stored in the vector $l$, where $l_i$ is the loudness of sound stream $i$. The loudness values are turned into a loudness ratio vector $r$, by taking logarithms, i.e.

$$r = 10log_{10}(l), \tag{8.1}$$

and the loudness balance, $b$, is calculated by offsetting the ratios to give a mean of zero,

$$b = r - r_{\mu}. \tag{8.2}$$

By balancing the loudness ratios about the zero mean, the loudness relationships between the component streams have been de-coupled from the overall loudness of the mix, defined as $m$. The 'mix' sound-stream is the summation of all component streams, and its short-term loudness is evaluated using the loudness model, and its overall loudness is calculated using Equations 7.4 and 7.9. The overall mix loudness, $m$, is an absolute loudness quantity, i.e. it is expressed in sones, unlike loudness ratios, which are expressed in decibels.

## 8.2 Automatic mixing case study

The objective mix feature used in Chapters 3 and 4 was a vector containing the relative levels of the component sounds. It is similar in form to the loudness balance in that it is a vector of relative quantities, but the normalisation was performed relative to the vocal level as opposed to the mean level. The reference mix was produced using a recorded version of the song, and if the user was an amateur musician, this would likely be a 'bedroom' mix, produced in the artist's home. The disparity in listening conditions between the bedroom and live mixes will likely cause differences in loudness balance, even if the relative levels are preserved.

This is demonstrated using the virtual live performance from Chapter 4 with scale factor, $d = 0$. The relative levels of the reference mix are retained, and its peak level is defined as 90 dBSPL[1]. The mix is optimised using the automatic mixing algorithm for the front-centre

[1] In the earlier automatic work the reference mix was defined using relative levels only, so an absolute level is defined here.

audience location, using the FOH loudspeakers only and no direct sound[2]. The loudness balance and overall mix loudness are extracted for both the reference and live mixes using the procedure outlined above, and are shown in Table 8.1.

| | Voice | Guitar | Bass | Kick | Snare | Hi-Hats | Cymbal | | Mix |
|---|---|---|---|---|---|---|---|---|---|
| $b_{ref}$ | 1.3 | -0.6 | -4.5 | -3.0 | 2.9 | 2.2 | 1.6 | $m_{ref}$ | 30.1 |
| $b_{live}$ | 0.0 | -1.4 | -2.0 | 0.5 | 2.3 | 0.1 | 0.0 | $m_{live}$ | 45.3 |
| $\delta b$ | -1.3 | –0.8 | +2.5 | +3.5 | -0.6 | -2.1 | -1.6 | | |

Table 8.1: The loudness balance and overall loudness of the reference 'bedroom' mix (peak level defined as 90 dBSPL) and the live mix, for the front centre audience location in the virtual live performance from Chapter 4 ($d = 0$), in which the relative levels have been reproduced exactly.

The loudness balance in the reference is different to that in the live mix, even though the objective mix features are the same, showing the importance of listening conditions when describing a mix, which cannot be accounted for when using objective or pseudo-perceptual features. The largest increases in relative loudness occur with the bass and drums. Both contain low frequency components, and as shown in Chapter 7, low frequency sounds get relatively louder with increased listening level.

Measuring the significance of loudness balance changes is an area of future work, but an indication as to whether they are noticeable can be obtained by considering the confidence intervals of the experimental data presented in Chapter 7 (summarised in Table 6.1). The mean interval for the simultaneous condition was 0.95 dB, which is smaller than the loudness balance changes in the: voice, bass, kick, hi-hats and cymbal. It is therefore likely that the differences would be noticeable. Within the existing live automatic mixing framework, it is simple to modify the system to operate on loudness balance, instead of relative levels.

## 8.3 A perceptual audio mixer

In this section a perceptual audio mixing device for recorded music is outlined, which replaces the traditional objective controls with perceptual controls, based on the loudness features that have been introduced. It can be applied to both live and recorded music, and in the case of live music, must incorporate the models developed in the early chapters of this thesis. For simplicity, it is demonstrated here for recorded music. Modern digital mixing systems are referred to as

[2]By using just one listening location and no direct sound the reference mix features can be recreated exactly.

*digital audio workstations* (DAWs), and operate on digital audio streams. They are analogous to the mixing console outlined in the general mix model (see Figure 2.1). In a multitrack setup, multiple digital audio streams are stored on separate 'tracks', and are mixed by manipulating the controls on traditional (music) signal processing devices.

The primary control parameter on a DAW is signal gain, which in analogue systems was controlled by logarithmic attenuating potentiometers, but which is now controlled by the DSP equivalent. Gain can be applied to each track via a fader, or to all tracks via the master fader. Although the loudness of a given sound stream can be modulated by changing the corresponding track gain, there is no direct mapping between gain and loudness. Furthermore, even if a nonlinear mapping were arrived at, it would be signal and listening-level dependent, and would provide no means to control masking interaction between sound-streams. In other words, relative positions of the faders on a mixing desk do not map directly to loudness ratios or loudness balance. The master fader on a DAW controls the output gain of the mix, but in most cases it does not control the overall listening level. Instead, a further gain control is employed somewhere in the monitoring signal chain (i.e. the loudspeaker amplifier).

The interface outlined here operates directly within the perceptual domain, and therefore provides simple, intuitive controls that account for the listening conditions. The controls consist of a loudness fader per track, from which loudness ratios can be evaluated directly to define the loudness balance, and a master loudness fader, which allows the overall loudness of the mix to be set directly.

### 8.3.1 Loudness estimation

The perceptual mixing system requires a calibration transfer function for the monitoring sound-system, $H(\omega)$, which converts audio-streams into sound-streams. This models the indirect signal path shown in the general mix model (see Figure 2.1), and for live applications would incorporate both direct and indirect signal paths. On a DAW this transfer function would convert dBFS into dBSPL. The audio stream in each track is therefore converted into a sound stream by applying the corresponding track and master fader gain values, and convolving with $H(\omega)$. It is from these sound-streams that the loudness balance $b$ and overall loudness $m$ are evaluated.

### 8.3.2 Optimisation strategy

The system uses an optimisation algorithm to find the gain applied to each track, which gives the required loudness balance and overall mix loudness. The subscripts $r$ and $t$ are introduced to identify the reference and target mix features respectively. The error variables $e_b$ and $e_m$ are given by,

$$e_b = b_r - b_t, \tag{8.3}$$

$$e_m = 10log_{10}(m_r/m_t). \tag{8.4}$$

The objective of the algorithm is to minimise these errors. The track and master fader gains, $g_t$ and $g_m$, are initially set to 0 dB, and at each iteration the errors are evaluated, and the gain required to correct them is estimated, using

$$g_{t,n} = g_{t,n-1} + e_{b,n-1} + e_{m,n-1}, \tag{8.5}$$

where $n$ identifies the current iteration. Equation 8.5 assumes a proportional relationship between the loudness and gain, so that a 5 dB loudness balance deficit would be corrected by adding 5 dB gain. Whilst the relationship is neither linear nor proportional, and does not take into account masking interactions, it is sufficiently well conditioned that over a number of iterations the solution converges. It is also worth noting that the master fader gain, $g_m$, is redundant because it has an equal effect on all tracks, and is therefore not included in this optimisation algorithm. However, $g_m$ is retained to allow the effect of master fader changes to be studied at a later stage.

### 8.3.3 Perceptual mixing case study

The perceptual mixer is demonstrated using a simple case study. Digital audio-streams are taken from a prerecorded mix containing eight instruments, and each stream is normalised to have a peak gain of 0 dBFS. The transfer function $H(\omega)$ defined in Section 3.1 is simplified to a scalar, and it is assumed that the sound re-enforcement system is calibrated such that 0 dBFS in the digital domain translates to 94 dBSPL in the pressure domain. The sound-streams therefore have peak levels of 94 dBSPL. Based on the heuristic models [Perez-Gonzales and Reiss, 2009a,b], the objective is make the component sound-streams equally loud, i.e. the reference loudness balance $b$ given by

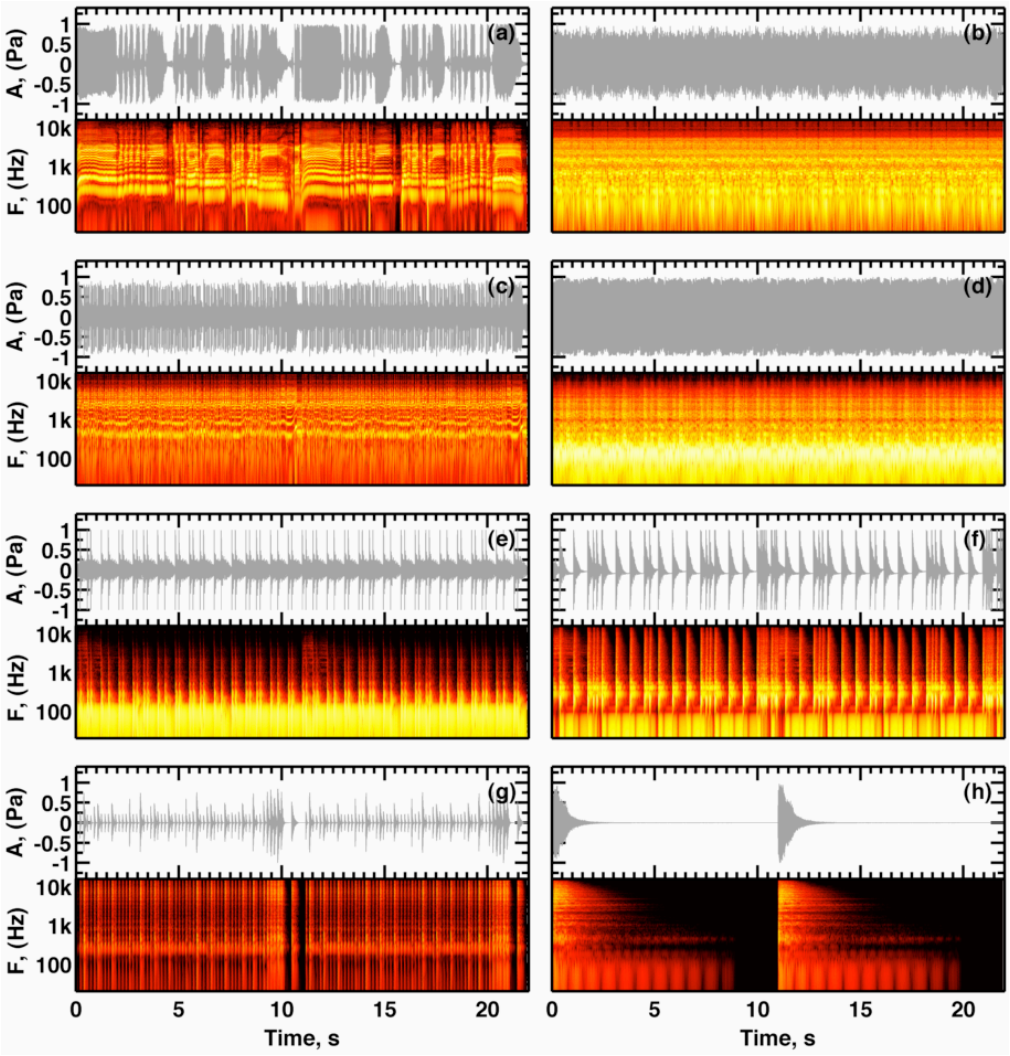


Figure 8.1: Waveforms and spectrograms of the sound signals used in this case study, corresponding to sources: (a)-(h) are voice, rhythm guitar, lead guitar, bass, kick drum, snare drum, hi-hats, cymbal.

$$b = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{8.6}$$

In addition, it is assumed that the loudness of the mix before optimisation, i.e the overall loudness when the eight sound-streams with peak levels of 94dBSPL are combined, must be preserved. The audio-streams were recorded from: (a) voice, (b) rhythm guitar, (c) lead guitar, (d) bass guitar, (e) kick drum, (f) snare drum, (g) hi-hats, and (h) cymbal. Waveforms and spectrograms of the corresponding sound-streams are shown in Figure 8.1.

The optimisation algorithm is initialised with track gains set to 0 dB, i.e. $g_{t,1} = 0$ for all tracks, as shown in the first row of Table 8.2. The next two rows show the reference loudness

balance, $b_r$, (equal loudness), and the starting target loudness balance, $b_{t,1}$. The optimisation algorithm is run, and the gain values at each iteration are shown in Figure 8.2. The loudness balance and gain values after 10 iterations are shown in the bottom two rows of Table 8.2. The loudness balance is realised with a tolerance of 0.01 dB, and the overall loudness $m$ is preserved within 0.6 sone, and would likely converge to the reference value if more iterations were allowed.

| | Voice | L. Guitar | R. Guitar | Bass | Kick | Snare | Hi-Hats | Cymbal | | Mix |
|---|---|---|---|---|---|---|---|---|---|---|
| $g_{t,1}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | $g_{m,1}$ | 0 |
| $b_r$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | $m_r$ | 77.5 |
| $b_{t,1}$ | 3.8 | 2.5 | 3.6 | 0.7 | -4.9 | -4.0 | -1.8 | 0.0 | $m_{t,0}$ | 77.5 |
| $b_{t,10}$ | -0.0 | 0.0 | -0.0 | 0.0 | 0.0 | 0.0 | -0.0 | -0.0 | $m_{t,10}$ | 76.9 |
| $g_{t,10}$ | -7.2 | -2.6 | -6.7 | 3.7 | 10.8 | 5.7 | 5.8 | 0.6 | $g_{m,10}$ | 0 |

Table 8.2: Gain values $g$, loudness balance $b$ and overall mix loudness $m$ for the equal loudness perceptual mixing case study. The subscripts $r$, $t$ and $n$ applied to $b$ and $m$ refer to the reference and target values, and the iteration number respectively.

The device outlined allows novice or expert audio engineers to perform mixing using perceptual controls. The user can directly manipulate the loudness balance and overall loudness, which are mapped to track gain parameters using an optimisation algorithm embedded in the device. The intuitive controls would be particularly useful for amateurs, who could specify, "...make the voice twice as loud as the guitars and the bass, and 3 times louder than the drums...". This could easily be achieved by moving the loudness faders to any positions that satisfy the required balance balance, i.e. voice on 6, guitars and bass on 3, and drums on 2. The intuitive controls could also take into account the interactions between sounds, which amateur audio engineers might find difficult to manage.

## 8.4 Method for estimating best practice for automatic mixing

A typical approach to automatic mixing is to apply heuristic models to the problem of setting control parameters of the traditional mixing device. These heuristic models discussed in Chapter 1 embody approximations to best practice, or else constitute assumptions about best practice (e.g., do not pan voice or low frequency sounds [Perez-Gonzales and Reiss, 2007, 2010], equal loudness probability among sound sources [Perez-Gonzales and Reiss, 2009a,b]). The concept of the perceptual mixer described in the previous section does not make any such assumptions. In fact, the process of defining the mix using perceptual controls is no more automated than for a traditional mixing desk. However, if a definition of best practice can be provided in terms of the perceptual controls, e.g. the loudness balance, there is no reason why it should not be input

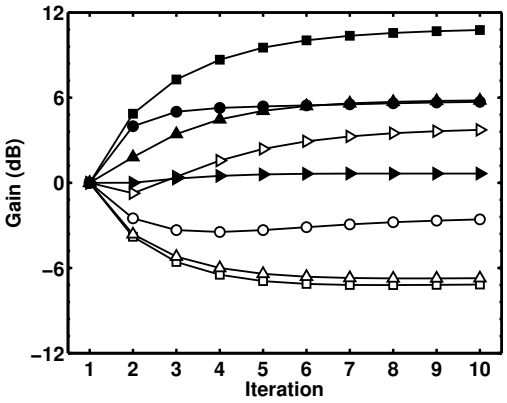


Figure 8.2: The gain settings applied at each iteration of the optimisation algorithm, the white markers: square, circles, vertical triangles and right pointing triangles correspond to voice, lead guitar, rhythm guitar and bass respectively. The shaded markers, following the same shape order, correspond to kick drum, snare drum, hi-hats and cymbal.

to the device in order to give fully automatic mixing functionality.

This section details a method of estimating best practice, where the system described is applied as an analytical tool. Five practicing audio engineers were assigned the task of producing a mix of the sound streams used in the previous section, using a traditional mixing interface. Random gain values were initially set on both the track and master faders, from -20 to 20 dB, with a maximum combined peak intensity of 112 dBSPL. Having obtained the track and master fader gain values from each participant, the model was used to estimate the loudness balance and overall mix loudness.

The loudness balance data, including mean and 95% confidence intervals, are shown in Figure 8.3. Clearly, all participants set the voice (instrument 1) to be louder than the other melodic instruments. The voice, and in particular the intelligibility of vocal lyrics, are important parts of popular music, so this result is not unexpected. The smallest variation in the loudness balance settings was for the bass guitar, and kick and snare drums (instruments 4 to 6). These instruments define the rhythmic structure of a song, and the results show that participants are in strong agreement as to how these should be balanced between one another, and with the melodic parts of the mix. The largest variation occurs with the rhythm and lead guitars. Figures 8.1 (b) and (c) show that these instruments occupy similar parts of the frequency spectrum, so there will inevitably be masking interactions that cause strong interdependency between the instrument loudness values, i.e. increasing the loudness of one will cause a decrease in the loudness of the other. This can be thought of as an instability in the loudness balance, because small changes in gain will cause

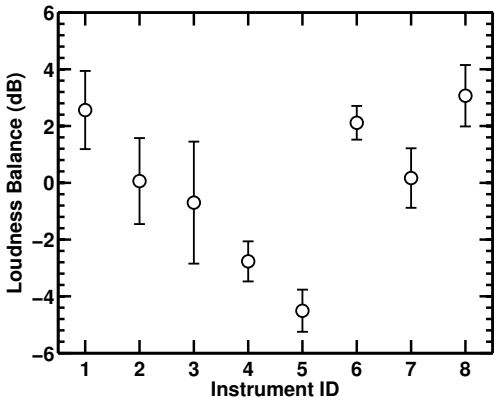


Figure 8.3: The best-practice loudness balance data extracted from mixes produced by practicing audio engineers. Shown are the mean values per instrument along with the 95% confidence intervals. The sound stream order from 1 to 8 are: voice, rhythm guitar, lead guitar, bass guitar, kick drum, snare drum, hi-hats, cymbal.

relatively large changes in loudness, and helps to explain the larger confidence intervals for these two instruments.

In principle, this data may be taken to embody best practice, as far as statistical confidence allows. The mean loudness balance becomes the new description of best practice for this combination of sound-streams, i.e. the equal loudness objective of Equation. 8.6 is replaced with,

$$b = \begin{bmatrix} 2.6 & 0.1 & -0.7 & -2.8 & -4.5 & 2.1 & 0.1 & 3.1 \end{bmatrix}. \tag{8.7}$$

This description may also be applied to alternate sound stream mixtures, to produce mixes according to this estimation of best practice. The data for this experiment do not support current heuristic models of best practice [Perez-Gonzales and Reiss, 2009a,b], which seek to produce equal loudness among sources. Based on the size of the 95% confidence intervals, it does not seem likely that this conclusion is caused by the relatively small population. However, the data presented here *is* a small sample, and larger scale testing is required to obtain reliable and transferable best practice mix descriptions. This should be done on many different mixes, within different genre, and where possible, with expert audio engineers. It is expected that best practice loudness balance will be genre and instrument-specific, so multiple best practice definitions will be needed. Detailed studies of best practice are an area of future work.

The overall loudness of the mixes produced by the participants had a mean of 47.4 sone and 95% confidence interval from 36.7 sone to 58.1 sone. One of the benefits of the loudness balance

feature is its de-coupling from the overall loudness, which made it possible to compare mixes produced at different levels. In addition, use of the model as an analytical tool has shown that participants tend to provide a common loudness balance, even though the overall mix loudness is varied.

The procedure outlined here describes how best practice loudness balance features can be evaluated. These features can be used as inputs to the perceptual audio mixer described in the previous section, to do fully automatic mixing, i.e. the embedded optimisation algorithm will determine the gain controls that realise the best practice loudness balance. As already mentioned, the descriptions may need to be genre and instrument-specific. Application of best practice would therefore involve a menu for the user to choose the best practice definition most relevant to them, e.g. jazz mix, rock mix, etc, but which can be adjusted further using the loudness balance controls.

## 8.5 A perceptual audio transmission format

The work in this thesis has sought to infer features of reference mixes onto target mixes. In Chapters 3 and 4 the target mixes were taken from a live performance, and the reference features were extracted from pre-produced reference mixes. In Sections 8.3 and 8.4 of this chapter, the target mix was a studio recording, and the reference features were input directly using perceptual controls, or were derived from mixing best practice experiments. The experiments in Chapter 7 showed that the perception of a reproduced mix will be affected by the listening conditions, and this applies to both live and recorded music. In this section, this concept is incorporated into a perceptual audio format, which encodes transmitted audio with reference sounds features, e.g. loudness balance, and which allows the errors upon reproduction to be estimated and corrected, for any listening conditions.

Figure 8.4 illustrates the audio production and transmission process, where the top part represents the studio in which the original mix is produced. The mixing engineer manipulates audio signals (denoted by $V(t)$ to show that they are voltage signals) or the digital equivalent, by listening to the sound streams $s_r(t)$, where the subscript $r$ identifies the reference mix. Once the mixing process is finished, the audio signal is transmitted either by radio, CD or digital download, and barring any data compression, the listener's audio signal is an identical copy. However, the listening conditions at reproduction will be different, i.e. different amplifier and loudspeaker

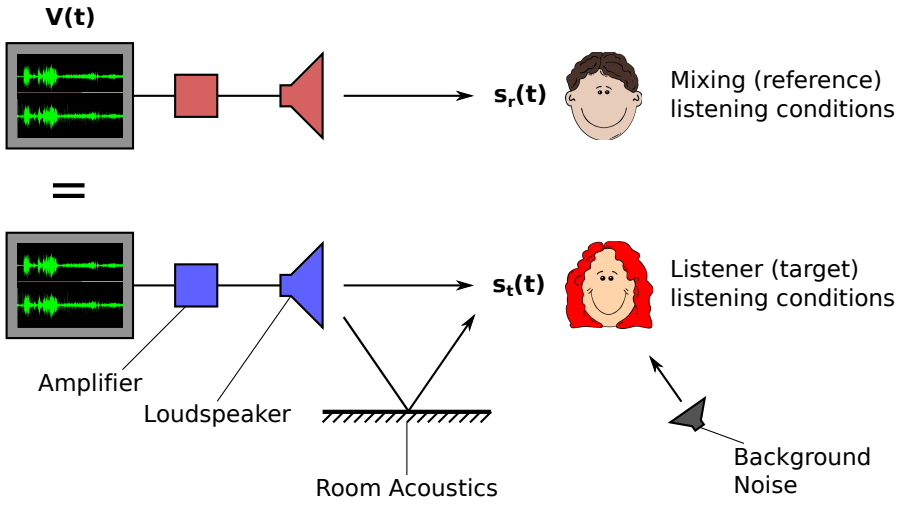


Figure 8.4: The audio mixing and transmission process, demonstrating the causes of differences in the sound-streams for reference and target reproduction.

characteristics, different room acoustics and the presence or not of background noise. This means that although the audio signal is identical, the sound-streams are different, and so the perception of the mix will change.

### 8.5.1 Sound Features

A sound can be described in terms of objective and perceptual features. To reiterate, objective features can be calculated directly from the acoustic signal e.g. the peak sound pressure level (SPL), whereas perceptual features are estimated based on the output of an auditory model, e.g. the loudness balance. The definition here is slightly different from that used in Chapter 6, in that the objective features must be calculated from a sound-stream as opposed to an audio-stream. The features of a sound-stream, $s(t)$, are stored in two vectors,

$$\eta = f(s), \tag{8.8}$$

$$\xi = g(s), \tag{8.9}$$

where the function $f(s)$ outputs objective features, and the function $g(s)$ outputs perceptual features (the time-dependence of $s(t)$ has been dropped for clarity).

### 8.5.2 Transmission Error

The errors in a transmission can be evaluated by comparing the objective and perceptual features of the reference and target sound-streams. The objective and perceptual error vectors are denoted by $E(\eta)$ and $E(\xi)$ respectively, where

$$E(\eta) = f(s_t) - f(s_r), \tag{8.10}$$

$$E(\xi) = g(s_t) - g(s_r). \tag{8.11}$$

For the live automatic mixing work in Chapters 3 and 4, the function $f(s)$ calculated the relative levels of the different instruments (see Equation 2.7), and the error metric $E(\eta)$ evaluated the difference between the reference and target mixes at all listener locations (see Equation 2.10). For the perceptual mixer outlined in this chapter, $g(s)$ calculated the loudness balance and the overall mix loudness, and the error $E(\xi)$ compared these with the reference values input using the perceptual controls (see Equation 8.4).

### 8.5.3 Definition of Format

The author of a transmitted sound, i.e. a mixing engineer, can make an informed decision as to the importance of different sound features by specifying the permitted errors in $E(p)$ and $E(\xi)$ and therefore the bounds within which the reproduced sound-streams are considered acceptable. These tolerances are stored and transmitted with the audio signal and are available to validate the reproduction. Formally, a production is said to be valid if

$$E(\eta) \leq E(\eta)_{Tol} \tag{8.12}$$

and

$$E(\xi) \leq E(\xi)_{Tol}, \tag{8.13}$$

where $E(p)_{Tol}$ and $E(\xi)_{Tol}$ are the allowable tolerances in objective and perceptual features respectively. This forms the basis of the proposed sound transmission format. It differs from existing transmission formats in that it encodes information about absolute pressure into the transmitted signals, and allows the related perceptual features to be evaluated.

### 8.5.4 Error correction

Optimisation algorithms, such as those already developed, can be employed to control signal processing tools that operate on the reproduced audio-streams, to reduce the objective and/or perceptual errors, depending on the format. For multi-stream features, like loudness balance, this would require the transmission of multitrack audio-streams.

### 8.5.5 Recorded music case study

The concept is demonstrated for recorded music, which people listen to in many different environments, from expensive loudspeakers in a quiet room to low quality headphones on a noisy underground train line. Due to the disparity in the target listening conditions it is unlikely that objective sound features can be recreated, so the format is defined using perceptual features. The chosen feature is the loudness balance, and the tolerance is set to 0.5 dB. In addition, the format is defined so that error correction, if required, does not alter the peak listening level[3].

A reference mix is produced using a digital audio workstation to represent the 'studio' recording. This mix is then reproduced at different listening levels and in different virtual environments. The listening conditions are: (i) living room; low level, room impulse response (RIR) applied representative of a small room, slight reduction in low frequency response to represent television loudspeakers, (ii) large venue; high level, RIR applied representative of a large, reverberant space, (iii) car; medium level, RIR applied representative of a typical in-car environment, road noise added.

The loudness balance is evaluated for the sound-streams in each listening condition, using the method outlined in Section 8.1, and are shown in Table 8.3. Also shown are the peak sound levels in dBSPL. The perceptual error, as defined by our format, is the difference in loudness balance between the studio mix and all other conditions, and is evaluated using Equation 8.13. These errors are reported in Table 8.4, and most are above the 0.5 dB tolerance defined in the format, so the transmission is invalid.

The optimisation algorithm outlined in Section 8.3 is used to minimise the transmission errors for each listening condition by adjusting track gain values, whilst constraining the overall peak level to its initial value. The track gain settings are shown in Table 8.5, which for all listening conditions reduced the error to less than 0.1 dB. Table 8.5 also shows that the peak levels are

[3]This allows the listener to set their listening level independently of the loudness balance correction. It would also be possible to constrain the overall loudness as in Section 8.3, rather than the level.

unchanged. The transmission is now valid based on the definition of the format.

| Condition | $SPL_{peak}$ | Voice | Guitar | Bass | Kick | Snare | Hi-Hats | Cymbal |
|---|---|---|---|---|---|---|---|---|
| Studio | 94 | 3.4 | -0.9 | -3.9 | -4.2 | 0.2 | 1.3 | 4.1 |
| Living Room | 88 | 4.6 | 2.4 | -3.5 | -5.6 | 2.0 | -1.8 | 2.0 |
| Large Venue | 106 | 2.0 | -1.1 | -2.7 | 2.4 | -2.9 | 0.4 | 1.9 |
| Car (with noise) | 100 | 4.2 | -0.6 | -4.9 | -8.1 | 1.4 | 2.8 | 5.3 |

Table 8.3: Feature extraction: the loudness balance of the mixes when reproduced under different listening conditions.

| Condition | $SPL_{peak}$ | Voice | Guitar | Bass | Kick | Snare | Hi-Hats | Cymbal |
|---|---|---|---|---|---|---|---|---|
| Living Room | 88.0 | 1.2 | 3.3 | 0.4 | -1.4 | 1.8 | -3.1 | -2.1 |
| Large Venue | 106.0 | -1.1 | -0.2 | 1.2 | 6.6 | -3.1 | -0.7 | -2.2 |
| Car (with noise) | 100.0 | 0.8 | 0.3 | -1.0 | -3.9 | 1.2 | 1.5 | 1.2 |

Table 8.4: Error reporting: the loudness ratio errors for the mixes at different listening conditions when compared to the original studio mix.

| Condition | $SPL_{peak}$ | Voice | Guitar | Bass | Kick | Snare | Hi-Hats | Cymbal |
|---|---|---|---|---|---|---|---|---|
| Living Room | 88.0 | -1.6 | -4.1 | 0.9 | 5.9 | -2.0 | 9.0 | 9.0 |
| Large Venue | 106.0 | 3.0 | 0.4 | -1.9 | -7.8 | 4.0 | 3.4 | 6.7 |
| Car (with noise) | 100.0 | -0.4 | 0.5 | 3.0 | 8.1 | -0.7 | -2.4 | -1.8 |

Table 8.5: Error correction: the signal gain applied to each track in the mix to correct the loudness balance errors and to preserve the peak level.

In this section a brief overview and demonstration of a perceptual audio transmission format has been given. The format uses absolute pressure information to extract objective and perceptual features from the reference and target mixes, from which transmission errors can be estimated and corrected. The format represents the extension of the live automatic mixing, and perceptual mixing work to encompass any audio transmission in general. The use of absolute pressure information, i.e. sound-streams instead of audio-streams, from which perceptual features can be evaluated, differentiates this from any existing audio format. Implementation of the format would improve the listening experience, because perception of a mix would be closer to that which was intended, and those who produce the content would have a means to quantify transmission errors introduced in different reproduction environments, and could even provide multiple mixes, with each mix tailored to suit specific listening conditions.

## 8.6 Summary

The loudness model for musical sound-streams, developed in the previous chapter, has been used to define the loudness balance; a new perceptual feature that describes a mix. It incorporates the effects of listening conditions on mix perception as well as the masking interactions between the component sounds. The availability of this feature, which is the first to be validated for music sounds, has many applications, some of which have been outlined here. These include: use of perceptual features for live automatic mixing, an audio mixing device that is operated using perceptual controls, a method to extract descriptions of mixing best practice to do fully automatic mixing, and a perceptual audio transmission format that enables the intended sound features to be recreated under any listening conditions. Validated perceptual features, of which the loudness balance is just one, enable us to define the objectives of music production process. If we have these features, using them to controlling music processing devices, though not always simple, is reduced to an engineering problem, that in most cases will be solvable. This contrasts with existing automatic mixing approaches, which implement a heuristic model, and use subjective evaluation to provide a qualitative assessment of automatic mixes, in comparison to manually produced mixes. In the next chapter general conclusions of this thesis are discussed, along with suggested extensions of the methodology used, and plans for future work.

# Chapter 9

# Conclusions and Future Work

---

This thesis has provided new methods and models that can be used to describe mixes, and to control associated mixing devices. The work is differentiable from other work in this field, generally termed automatic mixing, because analysis has been performed on acoustic signals of absolute pressure, as opposed to unscaled digital-audio signals; and the perceptual features developed have been validated for musical sounds. In this chapter, general conclusions are discussed, and areas for future work are briefly outlined.

## 9.1 The objectives of this thesis

In Chapter 1, a number of objectives were outlined, and are now discussed in turn.

### 9.1.1 A model of the mixing process

A general model of the mixing process was outlined in Chapter 2. This model included the direct signal paths for live acoustic sources, and the indirect signal path via the mixing console and loudspeakers. This model can be used for all mixing scenarios, as well as for reproduced music, to describe a mix in terms of its component acoustic signals. Areas of the model were developed further for live music applications, including models of source dispersion and room acoustic effects. Although some simplifying assumptions were made, these were shown to be related to the modelling of room impulse responses, and can easily be replaced with more advanced models or measurements.

### 9.1.2 Robust algorithms for live automatic mixing

The algorithms developed in Chapters 2 to 5 enable all practical issues surrounding live music production to be controlled automatically. It was necessary to split the optimisation algorithm into two stages in order to provide a robust and reliable solution. The main difficulties associated with live mixing, i.e. the direct sound, mix coupling, and acoustic feedback, were shown to be more significant in smaller venues, and it is in such places where amateur engineers, or musicians are most likely to attempt live mixing, making the automatic mixing system a very useful tool. Furthermore, the system would benefit more experienced engineers, because it could take on the functional burden of the mixing task, allowing them to concentrate on the more creative aspects of mixing. The complexity of the live mixing task has been reduced to the same level as for recorded mixing.

### 9.1.3 The effect of listening conditions on loudness perception

The psychophysical method of loudness ratio estimation was applied to musical sound-streams in Chapter 6. The estimated loudness ratios were shown to vary significantly with listening level for most pairs of sounds. From this it can be concluded that both objective, and pseudo-perceptual features, are unsuitable for describing a mix, unless the listening conditions are invariant.

### 9.1.4 To provide a validate loudness feature

The experimental data obtained in Chapter 6 was modelled in Chapter 7, by extending the existing loudness model of Glasberg and Moore [2002] for time-varying sounds. This introduced the dynamic sound stream bias (DSSB) coefficient, which weights the contribution of the mean and peak loudness time-function to the overall impression of loudness, and which is a function of the sound-stream crest factor. This model provided a better fit to the experimental data when compared to the state of the art. The loudness ratios were converted into the loudness balance, which is the first validated, truly perceptual mix feature, that can be used in automatic mixing algorithms.

### 9.1.5 To incorporate loudness features into automatic mixing systems

The loudness balance feature was used to demonstrate the perceptual errors that are introduced when live automatic mixing is done using objective sound features. Whilst not incorporated

explicitly, it would be a simple matter to substitute the relative level mix description, with the loudness balance, to enable live automatic mixing to be based on perceptual features.

A new, perceptual mixing system was outlined, that replaces the existing gain controls with loudness controls, enabling the user to manipulate perceptual sound features directly. It was demonstrated for recorded mixing, but by incorporating the full live mixing model, it can also be applied to live mixing. This system represents a step away from the existing automatic mixing approach [Barchiesi and Reiss, 2010, Perez-Gonzales and Reiss, 2009a,b, Terrell and Reiss, 2009], and is more akin to the re-parameterised audio effects outlined in Section 1.1.3. However, rather than re-paramaterising a single audio effect, the entire mixing system has been replaced with perceptual controls. Furthermore, the use of this system as an analytical tool provides a method of directly evaluating best practice, that can replace prior heuristic models.

### 9.1.6 To provide a framework for automatic mixing

The field of automatic mixing for music is in its infancy. It was arguably led by Perez-Gonzales and Reiss [2007, 2008, 2009a,b], who used heuristic models of mixing best-practice based on pseudo-perceptual loudness features. Their contribution cannot be overstated, particularly in the very early years of their research, but there are limits as to what can be achieved with their approach. Heuristic models must be tested subjectively, and this is generally done through qualitative comparisons of automatic and manually produced mixes. Such tests produce qualitative assessments of the automatic mix, for example, it is equally good compared to the manual mix. The heuristic models are described using sound features, but if the sound features have not been validated for musical sounds,[1] there will be errors in the descriptions. These errors cloud the conclusions of subjective tests, because there is uncertainty as to the features that made the automatic mix sound ‘good’ or ‘bad’. Furthermore, because the errors are likely to be sound-specific, a heuristic model that performs well on one mix, cannot be assumed to do so on another.

An alternative approach has been outlined in this thesis. It is centred on validated, perceptual features, e.g. the loudness balance, which are derived using rigorous psychophysical test methods. The availability of these features means that mixing best practice can be evaluated directly from manually produced mixes, and can be defined statistically. In order to do automatic mixing, the best practice mix description (which may be genre-specific) is input to the perceptual mixer, and can be thought of as a ‘mix preset’. Following from this, automatic mixing becomes a subset

[1] It has been demonstrated that pseudo-perceptual features cannot fully describe perceived loudness.

of perceptual mixing, with predefined mixes based on studies of best practice. The application of psychophysical methods to derive new mix features is the future of automatic mixing, because once we are able to describe perceptual features, it is relatively simple to collect best practice data, and manipulate signal processing devices to produce best practice mixes. It is therefore suggested that 'automatic' mixing might be re-branded as 'perceptual' mixing.

A number of assumptions were made in Chapter 2 in defining the engineer's role as an optimisation problem. Whilst focused on live mixing at the time, these assumptions provide a framework to do perceptual mixing. They are repeated here, and have been adapted slightly to apply to all forms of mixing.

1. The mixes can be modelled accurately, i.e. if the instrument signals and mixing console parameter settings are known, it is possible to evaluate the mix at any location as a set of acoustic signals.

2. The mix at a given location is *defined* as the linear superposition of multiple acoustic signals at that location, and a particular mix can be *described* using a set of *validated* perceptual features extracted from the acoustic signals from which it is composed.

3. The set of features that describe the reference mix, i.e. the objective of the mixing task, is available.

The first assumption ensures that the mix is available as a set of acoustic signals, the second states the need for a validate perceptual feature to describe the mixing process, and the third requires that a reference mix, or a set reference of features, are available to describe the objective. In the case of fully-automatic mixing, these can be derived from studies of best practice, and in the case of perceptual mixing tools, the reference features are input directly by the user, using a perceptual control interface.

## 9.2 Generalisation to other features

The loudness balance feature has been developed and validated in this thesis. It was chosen because loudness theory and models are mature and well understood, and because balancing the loudness of sounds within a mix is a critical part of the music production process. However, there are many other features related to, for example, spectral or temporal properties, that can be explored and defined, and this is an area for future work. New features can be drawn from

existing psychoacoustic research, which is vast, but which is limited with respect to music, and in particular mixes. This appears to be a very fertile field, because the psychophysical methods used in classical psychoacoustic studies have been shown to be applicable to musical sound-streams (see Chapter 7). The introduction of the DSSB coefficient provided a means to estimate the overall loudness impression of a musical sound-stream. The conversion of a feature time-function into a single quantity will likely be a common theme in the development of new features. The approach taken here, i.e. to weight the mean and peak of the time-function, is a simple model. Although it fits the data well, it is likely that more sophisticated models, which incorporate additional parameters, may provide a better representation of the feature time-functions.

## 9.3 Stream segregation

It was shown in Chapter 7 that participants are able to segregate sound-streams within a mix, and estimate their relative loudness quantities. A brief discussion of stream segregation was provided in relation to auditory scene analysis, but the process of stream segregation occurs when evaluating a mix, and therefore this requires further consideration.

The experiment in Chapter 7 showed that participants *can* segregate four simultaneous sound-streams and reliably estimate their loudness ratios. But does the loudness balance truly relate to their perception of the mix? Loudness balance is arguably a perceptual feature of a mix, but when we listen to a mix we listen to it as a whole, and do not (necessarily) attend to specific components, as was required in the experiment. The situation becomes more complex when the number of streams is increased. When listening to an orchestra it may be impossible for an untrained listener to segregate individual instruments, and it is more likely that they will group the instruments into sections, e.g. strings, which is arguably more useful, and which reflects the grouping used by the conductor.

A further aspect relating to stream segregation is that parts of the mixing process seek to combine streams from multiple acoustic sources into a signal sound. The temporal characteristics of a sound is one of the salient physical properties that promotes stream segregation [Moore and Gockel, 2002]. Dynamic audio effects, for example compressors, modify the temporal characteristics, and are often applied to groups of sounds, e.g. group compression on bass guitar and kick drum, and multi-band 'mastering' compression on a mix. When applied to groups of sounds, common temporal changes are inferred onto all, which will suppress stream segregation.

In these instances, the compressor is deliberately used to merge sound-streams, and has led to compressors being described as a 'glue' that binds components of a mix together. Therefore, a complete model and description of a mix cannot only consider the sound-streams in isolation, but must also include features of the mix as a whole.

For these reasons, current automatic mixing, or perceptual mixing work, would be more accurately described using the term 'sound balancing', as opposed to 'mixing'. Mixing implies blending of the components into a single sound, whereas balancing implies that the sounds are separate, and that their features must be defined with respect to one another, e.g. their loudness balance. The mixing process arguably includes a great deal of sound balancing, and it is the opinion of the author that it is a fundamental part of a mix, but it is not a complete mix description, so the distinction should be made.

## 9.4 Extensions to other fields of research

A further conclusion of this work, is that our perception of music is highly influenced by the listening conditions, and in particular the listening level. There are many areas within music technology research that could incorporate this, not least MIR, a large part of which seeks to provide perceptual features of music. At present, MIR features are almost exclusively extracted from audio-streams, because the listening conditions of the end user are not known. However, the perceptual audio format described in Section 8.5, which is outlined in more detail by Terrell et al. [2012] would make this possible. In addition, recent work which the author contributed to, showed that playlist recommendation systems, a stalwart MIR application, *are* dependent upon listening level, if based upon equivalent features to MFCCs that are calculated using an auditory model. It is therefore plausible that a new branch of MIR could be seeded from this work that makes use of perceptual, as opposed to pseudo-perceptual features, and which operates on sound-streams as opposed to audio-streams. This would be an important advancement in MIR, and also applies to other areas of music technology, because, to borrow a phrase from a close friend and colleague[2], we listen to **sounds**, not **signals**.

[2]Andrew Simpson: andy.simpson@eecs.qmul.uk.ac

# Bibliography

R. Aibara, J. T. Welsh, P. Sunil, and R. L. Goode. Human middle-ear sound transfer function and cochlear input impedance. *Hearing Research*, 152(100–109), 2001.

C. Alain, S. R. Arnott, S. Hevenor, S. Graham, and C. L. Grady. "what" and "where" in the human auditory system. *Proceedings of the National Academy of Science U.S.A.*, 98(21): 12301–12306, 2001.

J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

S. Anstis and S. Saida. Adaptation to auditory streaming of frequency-modulated tones. *The Journal of Experimental Psychology: Human Perception and Performance*, 11:257–271, 1985.

B. Bank. Perceptually motivated audio equalization using fixed-pole parallel second-order filters. *Signal Processing Letters, IEEE*, 15:477 –480, 2008.

D. Barchiesi and J. D. Reiss. Automatic target mixing using least-squares optimization of gains and equalization settings. *Proceedings of the 12th International Conference on Digital Audio Effects*, pages 7–14, 2009.

D. Barchiesi and J. D. Reiss. Reverse engineering of a mix. *The Journal of the Audio Engineering Society*, 58(7/8):563–576, 2010.

E. G. Bard, D. Roberton, and A. Sorace. Magnitude estimation of linguistic acceptability. *The Journal of the Linguistic Society of America*, 72(1):32–68, 1996.

H. Bauch. Die bedeutung der frequenczgruppe fur die lautheit von klangen. *Acustica*, 6:40–45, 1956.

L. L. Beranek. Acoustics. *New York: McGraw-Hill*, pages 183–207, 1954.

A. S. Bregman. Auditory streaming is cumulative. *The Journal of Experimental Psychology: Human Perception and Performance*, 4:380–387, 1978.

A. S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, Cambridge, MA, 1990.

R. Buckleinm. The audibility of frequency response irregularities. *The Journal of the Audio Engineering Society*, 1981:126–131, 29.

S. Buus, M. Florentine, and T. Poulsen. Temporal integration of loudness, loudness discrimination, and the form of the loudness function. *The Journal of the Audio Engineering Society*, 101(2):669–680, 1997.

K. B. Christensen. A generalization of the biquadratic parametric equalizer. *Proceedings of the 115th International Convention of the Audio Engineering Society*, 2003.

S.B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):357–366, 1980.

G. de Vries and G. van Beuningen. A digital control unit for loudspeaker array. *Proceedings of the 96th International Convention of the Audio Engineering Society*, 1994.

D. Dugan. Automatic microphone mixing. *Proceedings of the 51st International Convention of the Audio Engineering Society*, 1975.

D. Dugan. Application of automatic mixing techniques to audio consoles. *Proceedings of the 87th International Convention of the Audio Engineering Society*, 1989.

A. Eronen and A. Klapuri. Musical instrument recognition using cepstral coefficients and temporal features. *IEEE International Conference on Acoustics, Speech and Signal Processing*, (753–756), 2000.

S. Feistel and W. Ahnert. The significance of phase data for the acoustic prediction of combinations of sound sources. *Proceedings of the 119th International Convention of the Audio Engineering Society*, 2005.

S. Feistel and W. Ahnert. Modeling of loudspeaker systems using high-resolution data. *The Journal of the Audio Engineering Society*, 55(7/8):571–597, 2007.

S. Feistel, A. Thompson, and W. Ahnert. Methods and limitations of line source simulation. *The Journal of the Audio Engineering Society*, 57(6):379–402, 2009.

H. Fletcher. Auditory patterns. *Review of Modern Physics*, 12, 1940.

H. Fletcher and W. A. Munson. Loudness, its definition, measurement and calculation. *The Journal of the Audio Engineering Society*, 5:82–108, 1933.

H. Fletcher and W. A. Munson. Relation between loudness and masking. *The Journal of the Audio Engineering Society*, 1937.

M. Florentine, S. Buus, and T. Poulsen. Temporal integration of loudness as a function of level. *The Journal of the Audio Engineering Society*, 99:1699–1644, 1996.

International Organization for Standardization. ISO 226:2003(e): Acoustics; normal equal-loudness-level contours. Technical report, International Organization for Standardization, 2003.

B. M. Gibbs and D. K. Jones. A simple image source method for calculating the distribution of sound pressure levels within an enclosure. *Acustica*, 26:24–32, 1972.

B. R. Glasberg and B. C. J. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(1-2):103–138, 1990.

B. R Glasberg and B. C. J. Moore. A model of loudness applicable to time-varying sounds. *The Journal of the Audio Engineering Society*, 50:331–342, 2002.

B. R. Glasberg and B. C. J. Moore. Development and evaluation of a model for predicting the audibility of time-varying sounds in the presence of background sounds. *The Journal of the Audio Engineering Society*, 53(10):906–918, 2005.

B. R. Glasberg, B. C. J. Moore, R. D. Patterson, and I. Nimmo-Smith. Dynamic range an assymetry of the auditory filter. *The Journal of the Audio Engineering Society*, 76(2):419–427, 1984.

R. Greenfield and M. J. Hawskford. Efficient filter design for loudspeaker equalization. *The Journal of the Audio Engineering Society*, 39(10):739–751, 1991.

D. D. Greenwood. Critical bandwidth and the frequency coordinates of the basilar membrane. *The Journal of the Audio Engineering Society*, 33:1344–1356, 1961.

N. R. Haywood and B. Roberts. Effects of inducer continuity on auditory stream segregation: Comparison of physical and perceived continuity in different contexts. *The Journal of the Audio Engineering Society*, 130:2917–2927, 2011.

R. P. Hellman. Perceived magnitude of two-tone-noise complexes: Loudness, annoyance, and noisiness. *The Journal of the Audio Engineering Society*, 77:1497–1504, 1985.

M. Huotilainen, I. Winkler, K. Alho, C. Escera, J. Virtanen, R. J. Ilmoniemi, I. P Jãäskelãinen, E. Pekkonen, and R. Nãätãnen. Combined mapping of human auditory EEG and MEG responses. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 108(4):370 – 379, 1998.

American National Standards Institute. ANSI s3.4-2007: Procedure for computation of loudness of steady sounds. Technical report, American National Standards Institute, 2007.

S. Julstrom and T. Tichy. Direction-sensitive gating: a new approach to automatic mixing. *Proceedings of the 73rd International Convention of the Audio Engineering Society*, 1976.

M. Karjalainen and J. Mourjopoulos. About room response equalization and dereverberation. *2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 183–186, 2005.

M. Karjalainen and T. Paatero. Equalization of loudspeaker and room responses using kautz filters: Direct least squares design. *European Association for Signal Processing Journal on Advances in Signal Processing*, 2007.

D. T. Kemp. Stimulated acoustic emission from within the human auditory system. *The Journal of the Audio Engineering Society*, 64(5):1386–1391, 1978.

B. A. Kolasinski. A framework for automatic mixing using timbral similarity measures and genetic optimization. *Proceedings of the 124th International Convention of the Audio Engineering Society*, 2008.

G. F. Kuhn. The pressure transformation from a diffuse sound field to a the external ear and to the body and head surface. *The Journal of the Audio Engineering Society*, 65(4):991–1000, 1979.

H. Kuttruff. Room acoustics. *London: Applied Science Publishers*, 1979.

O. Lartillot, P. Toiviainen, and T. Eerola. A matlab toolbox for music information retrieval. In *Data Analysis, Machine Learning and Applications*, pages 261–268. Springer Berlin Heidelberg, 2008.

H. Levitt. Transformed up-down methods in psychoacoustics. *The Journal of the Audio Engineering Society*, 49(2):467–477, 1971.

B. Logan. Mel frequency cepstral coefficients for music modeling. *Proceedings of the International Symposium on Music Information Retrieval*, 2000.

S. Mansbridge, S. Finn, and J. D. Reiss. Implementation and evaluation of autonomous multitrack fader control. *Proceedings of the 132nd Audio Engineering Society Convention*, 2012.

B. McCarthy. Sound systems: Design and optimization. *Oxford: Elsevier Science Publishers Ltd*, 2007.

D. G. Meyer. Development of a model for loudspeaker dispersion simulation. *Proceedings of the 72nd International Convention of the Audio Engineering Society*, 1982.

D. G. Meyer. Computer simulation of loudspeaker dispersion. *The Journal of the Audio Engineering Society*, 32(5), 1984a.

D. G. Meyer. Digital control of loudspeaker array directivity. *The Journal of the Audio Engineering Society*, 32(10):747–754, 1984b.

Meyersound. UPA-1P datasheet. URL `http://www.meyersound.com/pdf/products/ultraseries/upa-1p_ds.pdf`.

Meyersound. Mapp Pnline Pro V.3.6.00, 2011. URL `http://www.meyersound.comproducts/mapponline/pro/`.

B. C. J. Moore. *An Introduction to the Physiology of Hearing*. Academic Press, 1997.

B. C. J. Moore and B. R. Glasberg. Formulae describing frequency selectivity as a function of frequency and level, and their use in calculateing excitation patterns. *Hearing Research*, 28 (2-3):209–225, 1987.

B. C. J. Moore and H. Gockel. Factors influencing sequential stream segregation. *Acta Acustica United with Acustica*, 88:320–333, 2002.

B. C. J. Moore, B. R. Glasberg, and T. Baer. A model for the prediction of thresholds, loudness, and partial loudness. *The Journal of the Audio Engineering Society*, 45:224–240, 1997.

B. C. J. Moore, S. Launer, D. Vickers, and T. Baer. Loudness of modulated sounds as a function of modulation rate, modulation depth, modulation waveform and overall level. *Psychophysical and Physiological Advances in Hearing*, 1998.

M. J. Morrell, C. A. Harte, and J.D. Reiss. Queen mary's "Media and Arts Technology Studios" Audio Sytem Design. Technical report, Audio Engineering Society, 2011.

J. Mourjopoulos. Digital equalization methods for audio systems. *Proceedings of the 84th International Convention of the Audio Engineering Society*, 1988.

J. Mourjopoulos. Digital equalization of room acoustics. *The Journal of the Audio Engineering Society*, 42(11):884–900, 1994.

S. T. Neely. A model of cochlear mechanics with outer hair cell motility. *The Journal of the Audio Engineering Society*, 94(1):137–146, 1993.

H. Nyquist. Regeneration theory. *Bell Systems Technical Journal*, 11:136–147, 1932.

A. J. Oxenham and C. J. Plack. A behavioural measure of basilar-membrane nonlinearity in listeners with normal and impaired hearing. *The Journal of the Audio Engineering Society*, 101(6):3666–3675, 1997.

T. Paatero and M. Karjalainen. Kautz filters and generalized frequency resolution: Theory and audio applications. *The Journal of the Audio Engineering Society*, 51(1/2):27–44, 2003.

E. Pampalk. *Computational models of music similarity and their application to music information retrieval*. PhD thesis, Johannes Kepler Universitat Linz, 2006.

R. D. Patterson. Auditory filter shapes. *The Journal of the Audio Engineering Society*, 55(4): 802–809, 1974.

R. D. Patterson. Auditory filter shapes derived with noise stimuli. *The Journal of the Audio Engineering Society*, 59(3):640–654, 1976.

R. D. Patterson and I. Nimmo-Smith. Off fequency listening and auditory-filter asymmetry. *The Journal of the Audio Engineering Society*, 67(1):229–245, 1980.

R. D. Patterson, I. Nimmo-Smith, and R. Milroy. The deterioration of hearing with age: frequency selectivity, the critical ratio, the audiogram, and speec threshold. *The Journal of the Audio Engineering Society*, 72(6):1788–1803, 1982.

E. Perez-Gonzales and J. D. Reiss. Automatic mixing: live downmixing stereo panner. *Proceedings of the 10th International Conference on Digital Audio Effects*, 2007.

E. Perez-Gonzales and J. D. Reiss. An automatic maximum gain normalization technique with applications to audio mixing. *Proceedings of the 124th International Convention of the Audio Engineering Society*, 2008.

E. Perez-Gonzales and J. D. Reiss. Automatic equalization of multi-channel audio using cross-adaptive methods. *Proceedings of the 129th International Convention of the Audio Engineering Society*, October 2009a.

E. Perez-Gonzales and J. D. Reiss. Automatic gain and fader control for live mixing. *The IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2009b.

E. Perez-Gonzales and J. D. Reiss. A real-time semi-autonomous audio panning system for music mixing. *European Association for Signal Processing Journal on Advances in Signal Processing*, 2010.

J. O. Pickles. *An Introduction to the Physiology of Hearing*. Academic Press, 2008.

S. Puria, W. T. Peake, and J. J. Rosowski. Sound-pressure measurements in the cochlear vestibule of human-cadaver ears. *The Journal of the Audio Engineering Society*, 101(5):2754–2770, 1997.

Z. Rafii and B. Pardo. Learning to control a reverberator using subjective perceptual descriptors. *Proceedings of the 10th International Society for Music Information Retrieval Conference*, 2009.

G. Ramos and J. L. López. Filter design method for loudspeaker equalization based on iir parametric filters. *The Journal of the Audio Engineering Society*, 54(12):1162–1178, 2006.

D. Reed. A preceptual assistant to do sound equalization. *Proceedings of the 5th International Conference on Intelligent User Interfaces*, 2000.

W. S. Rhode. Observations of the vibration of the basilar membrane in squirrel monkeys using the mossbauer technique. *The Journal of the Audio Engineering Society*, 49:1218–1231, 1971.

L. Robles, M. A. Ruggero, and N. C. Rich. Basilar membrane mechanics at the based of the chinchilla cochlea. input-output functions, tuning curves, and response phases. *The Journal of the Audio Engineering Society*, 80:1364–1374, 1986.

A. T. Sabin and B. Pardo. A method for rapid personalization of audio equalization parameters rapid personalization of audio equalization parameters. *Proceedings of the 17th ACM Conference on Multimedia*, pages 769–772, 2009a.

A. T. Sabin and B. Pardo. 2deq: an intuitive audio equalizer. *Proceeding of the 7th ACM Conference on Creativity and Cognition*, pages 435–436, 2009b.

E. A. G. Shaw. Transformation of sound pressure level from the free field to the eardrum in the horizonatl plane. *The Journal of the Audio Engineering Society*, 56(6):1848–1861, 1974.

J. O. Smith. Bark and erb bilinear transforms. *IEEE Transactions on Speech and Audio Processing*, 7(6):697–708, 1999.

H. Staffeldt and A. Thompson. Line array performance at mid and high frequencies. *Proceedings of the 117th International Convention of the Audio Engineering Society*, 2004.

S. S. Stevens. On the psychophysical law. *Psychological Review*, 64(3):152–181, 1957.

S. S. Stevens. *Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects*. New York: Wiley, 1975.

M. J. Terrell and J. D. Reiss. Automatic monitor mixing for live musical performance. *The Journal of the Audio Engineering Society*, 57(11):927–936, November 2009.

M. J. Terrell, J. D. Reiss, and M. Sandler. Automatic noise gate settings for drum recordings containing bleed from secondary sources. *European Association for Signal Processing Journal on Advances in Signal Processing*, 2010.

M. J. Terrell, A. J. R. Simpson, and M. Sandler. Sounds not signals: A perceptual audio format. Technical Report 52, Audio Engineering Society, 2012.

A. Thompson. Line array splay angle optimisation. *Proceedings of the Institute of Acoustics*, 28, 2006.

A. Thompson. Improved methods for controlling touring loudspeaker arrays. *Proceedings of the 127th International Convention of the Audio Engineering Society*, 2009.

A. Thompson, J. Baird, and B. Webb. Numerically optimised touring loudspeaker arrays - practical applications. *Proceedings of the 131st International Convention of the Audio Engineering Society*, 2011.

Tontechnik-Rechner. Absoprtion Coefficients, 2012. URL `http://www.sengpielaudio.com/calculator-RT60Coeff.htm`.

F. E. Toole. Loudspeaker measurements and their relationsihp to listener preference. *The Journal of the Audio Engineering Society*, 34(1):227–235, 1986.

F. E. Toole and S. E Olive. The modification of timbre by resonances: perception and measurement. *The Journal of the Audio Engineering Society*, 36(3):122–142, 1988.

G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(293–301), 2002.

European Broadcast Union. R-128: Loudness normalisation and permitted maximum level of audio signals. Technical report, EBU, 2011.

M. S. Ureda. J and spiral line arrays. *Proceedings of the 111th International Convention of the Audio Engineering Society*, 2001.

J. van der Werff. Design and implementation of a sound column with exceptional properties. *Proceedings of the 96th International Convention of the Audio Engineering Society*, 1994.

T. van Waterschoot and M. Moonen. Fifty years of acoustic feedback control: State of the art and future challenges. *Proceedings of the IEEE*, 99(2):288 –327, feb. 2011.

V. Verfaille, U. Zoelzer, and D. Arfib. Adaptive digital audio effects (a-dafx): A new class of sound transformations. *IEEE Transactions on Audio, Speech and Language Processing*, 14 (5):1817–1831, 2006.

J. H. Ward. Hierarchical grouping to optimize an objective function. *The Journal of the American Statistical Association*, 58:236–244, 1963.

B. Webb and J. Baird. Advances in line array technology for live sound. *Proceedings of the 18th International Conference of the Audio Engineering Society*, 2003.

D. L. Weber. Growth of masking and the auditory filter. *The Journal of the Audio Engineering Society*, 62(2):424–429, 1977.

I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Oxford: Elsevier Science Publishers Ltd, 2005.

C. Zhang and F. G. Zeng. Loudness of dynamic stimuli in acoustic and electric hearing. *The Journal of the Audio Engineering Society*, 102:2925–2934, 1997.

E. Zwicker. Procedure for calculating loudness of temporally variable sounds. *The Journal of the Audio Engineering Society*, 62(3):675–682, 1977.

E. Zwicker and H. Fastl. *Psychoacoustic - Facts and Models*. Springer-Verlag, Berlin, 1990.

E. Zwicker and B. Scharf. A model of loudness summation. *Psychological Review*, 72(1):3–26, 1965.

E. Zwicker, G. Flottorp, and S. S. Stevens. Critical band width in loudness summation. *The Journal of the Audio Engineering Society*, 29(5):548–557, 1957.